



*The Proceedings of the 7th Annual International
Conference on Digital Government Research*

*San Diego, California
U.S.A.*

May 21-24, 2006

Published by the Digital Government Research Center
© 2006, Digital Government Research Center

For information on this and other DGRC publications, contact info@dgrc.org

CONFERENCE

CONFERENCE CO-CHAIRS:

José A. B. Fortes, University of Florida
Ann Macintosh, Napier University

PROGRAM CO-CHAIRS:

Judith B. Cushing, The Evergreen State College
Theresa Pardo, Center for Technology in Government and SUNY Albany

WORKSHOP AND TUTORIAL CHAIR:

Stuart Shulman, University of Pittsburgh

BIRDS-OF-A-FEATHER CHAIR:

J. Ramón Gil-Garcia, Center for Technology in Government, University Albany/SUNY

SYSTEM DEMONSTRATION AND POSTER SESSION CO-CHAIRS:

Eduard Hovy, Digital Government Research Center, University of Southern California

Peggy Agouris, University of Maine

FINANCE CHAIR:

Yigal Arens, Digital Government Research Center, University of Southern California

PUBLICITY CHAIR:

Chrystol Koempel, Digital Government Research Center, University of Southern California

STUDENT CHAIR:

Natalie Helbig, Center for Technology in Government, University Albany/SUNY

GOVERNMENT LIAISON:

Lawrence Brandt, National Science Foundation

CONFERENCE COMMITTEE MEMBER:

Hans (Jochen) Scholl, University of Washington

COMMUNICATIONS AND OUTREACH CHAIR:

Valerie Gregg, Digital Government Research Center, University of Southern California

REGISTRATION AND LOCAL ARRANGEMENTS FOR DG.O 2006:

Priscilla Rasmussen, Academic and Research Conference Services (ARCS)

A Special Thank You to Our Sponsor

dg.o 2006 is generously supported by a grant from:



Grant No. IIS-0518024

Friends and colleagues,

Allow me to welcome you to dg.o 2006, our seventh annual international conference for multi-disciplinary and cross-sector teams in digital government (DG) research. I would like to extend a special welcome to new digital government principal investigators (PIs) and government partners who are here for their first dg.o event. I can promise you a stimulating three days, as those who have been here before will testify.

I want to thank the people who have made this event possible:

- José A. B. Fortes and Ann Macintosh, conference co-chairs;
- Judy Cushing and Theresa Pardo, program co-chairs;
- Stuart W. Shulman, J. Ramón Gil-Garcia, Eduard Hovy, Peggy Agouris, Natalie Helbig, Valerie Gregg and Jochen Scholl; conference committee members;
- Yigal Arens, Chrystol Koempel, and Priscilla Rasmussen who coordinated financial and logistical aspects of the conference.

The up-front work of this group of experts will enable you to experience the breadth and depth of DG research that has been accomplished by our diverse and multi-disciplinary community of researchers, students, and government managers. Also, I would like to thank those of you who submitted white papers and other proposals; whether or not your paper or proposal was accepted, the conference benefits from your creativity and intellectual contributions.

Many of you know that the Information and Intelligent Systems Division, including the Digital Government Program, has been reorganized with a goal to simplify the submission of proposals, improve the review process, lower barriers between programs within the division and allow funding of truly excellent larger proposals. By the time this conference occurs, a new division-wide call for proposals will have been issued, which I will be prepared to discuss. For now, I'm happy to report that the enthusiasm is reflected within the restructured division and the CISE Directorate, as the digital government program continues to grow. DG is seen as an important generator of exciting proposals and as a critical application area for information, telecommunication and computer science research (along with related social and behavioral sciences).

Our community maintains connections throughout the year thanks to the hard work of Chrystol Koempel, our Public Affairs/Media Officer, who is responsible for the DG website and the monthly dg.Online newsletter. Please coordinate with Chrystol to ensure a strong venue for your work. This year as in the past, your response to our call for Project Highlights has allowed me to respond effectively to the requirements of the Federal Government Performance and Results Act (GPRA). GPRA is a critical activity to inform the National Science Foundation (NSF) management of your exciting results. I appreciate your timeliness and responsiveness in the GPRA process.

One of the real values of this conference will happen when you take the opportunity to reach to those outside your normal circle of colleagues; you will find potential new colleagues reaching back in response. Though initial cross-disciplinary discussions take some time, I think you will find them worthwhile; and, NSF is always interested in those kinds of new partnerships.

I hope to meet and say, "Hello" to all of you during the conference; please find me and make sure that happens.

Your NSF Program Manager for Digital Government Research,
Lawrence E. Brandt

Welcome to dg.o 2006!

It is our pleasure to welcome you to dg.o 2006, The Seventh Annual International Conference on Digital Government Research. Dg.o 2006 is unique in bringing together computer and social science researchers, government officials and representatives of industry to investigate how information and communication technologies can be used to make government information, processes and services better and more conveniently and securely available to all stakeholders.

The conference has grown considerably over the last six years - beginning as a national conference for U.S. researchers to a truly international conference attracting participants worldwide. Last year saw the successful introduction of International Research Workshops and Tutorials; this year we offer these again on a number of topics of shared interest to digital government researchers in different countries.

We are also extremely fortunate to have three distinguished keynote speakers. Dr. Fran Berman, San Diego Supercomputer Center Professor and High Performance Computing Endowed Chair at UC San Diego, will deliver a conference keynote on Monday morning focusing on the past one hundred years of data and the challenges facing data management in the foreseeable future. On Tuesday morning, participants of dg.o 2006 and ISI 2006 will come together to hear from Dr. John R. Phillips the Chief Scientist for the Central Intelligence Agency. Dr. Philips will focus on information technology research for national security. Then on Tuesday afternoon, Kimberly Nelson, Executive Director eGovernment at Microsoft Corporation, will consider the importance of academic and business partnerships that have the potential to enhance digital government research.

Dg.o 2006 is a forum for the presentation and discussion of interdisciplinary research on digital government and its applications in diverse domains. In addition to refereed research papers, we will also have panel discussions on key issues, system demonstrations, poster sessions, tutorials, and Birds-of-a-Feather roundtables. This year, 58 research paper submissions were received, 51 of which were complete and sent to reviewers. Of those 51, 20 were accepted as research papers for the conference. In addition to research papers, 60 project highlights will be presented at the conference, some as posters or system demonstrations and some as short research presentations. In addition we have 45 research posters, and 8 system demonstrations.

Above all the conference aims to bring together researchers and developers from a wide range of academic, research, governmental, and industrial backgrounds. Delegates will have the opportunity to discuss their ideas and problems with researchers working at the leading edge of digital government developments and with government officials and technological developers with experience in the practical issues of getting systems into operation. We have arranged for breaks and common lunches and evening sessions to facilitate collaboration in the development of great new ideas. We expect to see the results of these at dg.o2007 and other future digital government conferences!

The conference is generously supported by a grant from the National Science Foundation, and is organized by the Digital Government Research Center. DGRC is a joint center of the USC Information Sciences Institute and Columbia University.

We acknowledge the insightful contributions of the dg.o 2006 Program Committee:

Peggy Agouris, University of Maine, USA
José Luis Ambite, University of Southern California, USA
Yigal Arens, University of Southern California, USA
Alexis Barabashev, Higher School of Economics, Russia
Chaitan Baru, University of California San Diego, USA
Alan Borning, University of Washington, USA
Shawn Bowers, University of California, San Diego, USA
Laura Bright, Portland State University, USA
Jamie Callan, Carnegie Mellon University, USA
Hsinchun Chen, University of Arizona, USA
Leslie Cheung, University of Southern California, USA
Anthony Cresswell, University at Albany/SUNY, USA
Judith Bayard Cushing, The Evergreen State College, USA
Sharon S. Dawes, University at Albany/SUNY, USA
Lois Delcambre, Portland State University, USA
José A. B. Fortes, University of Florida, USA
Jane Fountain, University of Massachusetts, Amherst, USA
Jagdish Gangolly, University at Albany/SUNY, USA
Jon Gant, Syracuse University, USA
J. Ramón Gil-García, University at Albany/SUNY, USA
Genevieve Giuliano, University of Southern California, USA
Mike Goodchild, University of California, Santa Barbara, USA
Åke Grönlund, Orebro University, Sweden
Andy Hamilton, University of Salford, UK
Teresa Harrison, University at Albany/SUNY, USA
Thomas Horan, Claremont Graduate Institute, USA
Eduard Hovy, University of Southern California, USA
Bill Howe, Portland State University, USA
Marijn Jansen, Delft University of Technology, The Netherlands
Alan Karr, National Institute of Statistical Sciences, USA
Bernhard Katzy, CeTIM, The Netherlands
Jay Kesan, University of Illinois, USA
Andrei Klimentko, Higher School of Economics, Russia
Gloria Lau, Stanford University, USA
Kincho Law, Stanford University, USA
Man-Sze Li, IC Focus Ltd, UK
Ann Macintosh, Napier University, UK
Gary Marchionini, University of North Carolina at Chapel Hill, USA
Enrico Nardelli, NESTOR - Univ. Roma 'Tor Vergata', Italy
Paul O'Connell, Iona College, USA

John O'Flaherty, The National Microelectronics Applications Centre Ltd., Ireland
Patrick Pantel, University of Southern California, USA
Theresa Pardo, University at Albany/SUNY, USA
Manuel Joao Pereira, INA-National Administration Institute-Technical Institute of Portugal, Portugal
Doncho Petkov, Eastern Connecticut State University, USA
H. Raghav Rao, University of Buffalo/SUNY, USA
Charles Rothwell, National Center for Health Statistics, USA
Hans (Jochen) Scholl, University of Washington, USA
Rajiv Shah, University of Illinois, USA
Stuart Shulman, University of Pittsburgh, USA
David Socha, University of Washington, USA
Anthony Stefanidis, University of Maine, USA
Laura Steinberg, George Washington University, USA
Jennifer Stromer-Galley, University at Albany/SUNY, USA
Luis Valadares Tavares, INA-National Administration Institute-Technical Institute of Portugal, Portugal
Giri Kumar Tayi, University at Albany/SUNY, USA
Fiona Thompson, University at Albany/SUNY, USA
Roland Traumuller, University of Linz, Austria
Vassilis Tsotras, University of California, Riverside, USA
Trond Undheim, European Commission, Belgium
Bienvenido Velez-Rivera, U of Puerto Rico Mayaguez, USA
Paul Waddell, University of Washington, USA
Maria Wimmer, University of Koblenz-Landau, Germany
Dawn Wright, Oregon State University, USA
Alexandros Xenakis, Napier University, UK
Yan Yang, University of Southern California, USA
James P. Zappen, Rensselaer Polytechnic Institute, USA
Junliang Zhang, University of North Carolina at Chapel Hill, USA

Student Program Committee

Phang Chee Wei, National University of Singapore, Singapore
Gabriel Puron Cid, University at Albany/SUNY, USA
Kristin Eickhorst, University of Maine, USA
Natalie Helbig, University at Albany/SUNY, USA
Soo-Min Kim, USC Information Sciences Institute, USA
Namhee Kwon, USC Information Sciences Institute, USA
Shuhua Liu, University of Washington, USA
Pee Loo Geok, National University of Singapore, Singapore
Alexander Schellong, Harvard University, USA
Kristene Unsworth, University of Washington, USA
Liang Zhou, USC Information Sciences Institute, USA

We thank Jim Costello (Center for Technology in Government, University at Albany/SUNY) for his excellent technical support for the automated paper submission and review management system.

Planning and arranging the conference, including the conference program, was truly a team effort. We hope you enjoy the conference!

José A. B. Fortes, Conference Co-Chair

Ann Macintosh, Conference Co-Chair

Judith B. Cushing, Program Co-Chair

Theresa A. Pardo, Program Co-Chair

Lawrence Brandt, Government Liaison

Stuart W. Shulman, University of Pittsburgh, Workshop and Tutorial Chair

J. Ramón Gil-García, Birds-of-a-Feather Chair

Eduard Hovy, System Demonstration and Poster Session Co-Chair

Peggy Agouris, System Demonstration and Poster Session Co-Chair

Hans (Jochen) Scholl, Conference Committee Member

Yigal Arens, Finance Chair

Natalie Helbig, Student Chair

Valerie Gregg, Communications and Outreach Chair

Chrystol Koempel, Publicity Chair

Priscilla Rasmussen, Registration and Local Arrangements for dg.o 2006

Erika Barragán-Nuñez, Administrative Support

Alma Nava, Administrative Support

TABLE OF CONTENTS

Conference Officers	iii
NSF Welcome	vii
Conference Committee Welcome	viii
Table of Contents	xii
Program at a Glance	xxi
Author Index	xxiii
INVITED TALKS	1
One Hundred Years of Data	3
Fran Berman, University of California San Diego, USA	
Academic and Business Partnerships to Enhance Digital Government Research	5
Kimberly Nelson, Microsoft Corporation, USA	
PANELS	7
SESSION 1C: Cyberstructure for Public Health: Digital Government Research Collaborators	9
<i>Moderator: Noshir Contractor, University of Illinois at Urbana-Champaign, USA</i>	
PLENARY: National Science Foundation Program Managers: Perspectives on Sustaining Digital Government Research	11
<i>Moderator: Valerie Gregg, University of Southern California, USA</i>	
SESSION 2C: Ecosystem Informatics for Decision-Makers	12
<i>Moderator: Tyrone Wilson, USGS National Biological Information Infrastructure, USA</i>	
SESSION 4B: Sustaining an International Digital Government Research Community	14
<i>Moderator: Sharon S. Dawes, University at Albany/SUNY, USA</i>	
RESEARCH PRESENTATIONS	17
SESSION 1A: DATA MINING	19
<i>Moderator: Jamie Callan, Carnegie Mellon University, USA</i>	
Spatiotemporal Analysis of 9-1-1 Call Stream Data ♦	21
Jasso, Hector; Hodgkiss, William; Fountain, Tony; Baru, Chaitan; Reich, Don; Warner, Kurt	
A Temporal Based Forensic Analysis of Electronic Communication Ω	23
Stolfo, Salvatore J.; Hershkop, Shlomo; Creamer, German	
Using Semantic Components to Facilitate Access to Domain-Specific Documents in Government Settings ♦	25
Price, Susan; Delcambre, Lois; Nielsen, Marianne Lykke; Tolle, Timothy; Luk, Vibeke; Weaver, Mathew	
SESSION 1B: VOTING	27
<i>Moderator: James P. Zappen, Rensselaer Polytechnic Institute, USA</i>	
Political E-Identity: Campaign Funding Data and Beyond Δ	29
Wolber, David	
A Project to Assess Voting Technology and Ballot Design Ω	38
Traugott, Michael W.; Hernson, Paul S.; Niemi, Richard G.	
E-voting in the 2005 Local Elections in Estonia and the broader impact for future e-voting projects Ω	40
Trechsel, Alexander H.; Breuer, Fabian	

SESSION 2A: INTEGRATED JUSTICE	35
<i>Moderator: Hans J. (Jochen) Scholl, University of Washington, USA</i>	
Research Issues Related to Exchanging Information from Heterogeneous Data Sources Ω	45
Swigger, Kathleen; Brazile, Robert	
Integrated Criminal Justice: ARJIS Case Study Ω	47
Sawyer, Steve; Tyworth, Michael	
COPLINK Center: Social Network Analysis and Identity Deception Detection for Law Enforcement and Homeland Security Intelligence and Security Informatics: A Crime Data Mining Approach to Developing Border Safe Research Ω	49
Chen, Hsinchun; Atabakhsh, Homa; Wang, Alan G.; Kaza, Siddharth; Tseng, Lu Chunju; Wang, Yuan; Joshi, Shailesh; Petersen, Tim; Violette, Chuck	
SESSION 2B: CITIZEN PARTICIPATION 1	51
<i>Moderator: Teresa Harrison, University at Albany/SUNY, USA</i>	
Should E-Government Design for Citizen Participation? Stealth Democracy and Deliberation Δ	53
Muhlberger, Peter	
Converting Online Public Legal Information into Knowledge: "ABC del Diritto" an Italian e-Government Citizen-oriented Service Δ	62
Biasiotti, Maria Angela; Nannucci, Roberta	
Web Portal Implementation to Support Public Participation in Transportation Decision Making Ω	67
Nyerges, Tim; Brooks, Terry; Jankowski, Piotr; Rutherford, G. Scott; Young, Rhonda	
SESSION 3A: CRISIS MANAGEMENT 1	69
<i>Moderator: José Fortes, University of Florida, USA</i>	
GeoCollaborative Crisis Management: Designing Technologies to Meet Real-World Needs Ω	71
MacEachren, Alan M.; Cai, Guoray; McNeese, Michael; Sharma, Rajeev; Fuhrmann, Sven	
Indexing and Searching Handwritten Medical Forms Ω	73
Govindaraju, Venu	
Digital Governance and Hotspot Geoinformatics for Monitoring, Etiology, Early Warning, and Management Around the World Ω	75
Patil, G.P.	
SESSION 3B: CITIZEN PARTICIPATION 2	77
<i>Moderator: Thomas Horan, Claremont Graduate Institute, USA</i>	
When Opinion Leaders Blog: New forms of citizen interaction Δ	79
Kavanaugh, Andrea; Zin, Than Than; Carroll, John M.; Schmitz, Joseph; Pérez-Quiñones, Manuel; Isenhour, Philip	
Bringing an Informed Public into Policy Debates through Online Deliberation: The Healthcare Dialogue project Ω	89
Price, Vincent; Cappella, Joseph N.	
Decoding Political Discourse Networks Ω	91
Kelly, John; Stark, David	
SESSION 3C: EMERGENT DESIGN	93
<i>Moderator: Alan Borning, University of Washington, USA</i>	
Water Models and Water Politics: Design, Deliberation, and Virtual Accountability Δ	95
Jackson, Steven	
Organic Development: A Top-Down and Bottom-Up Approach to Design of Public Sector Δ Information Systems	105
Tyworth, Michael; Sawyer, Steve	

SESSION 4A: CRISIS MANAGEMENT 2	113
<i>Moderator: Anthony Cresswell, University at Albany/SUNY, USA</i>	
TIME-CRITICAL INFORMATION SERVICES: Analysis and Workshop Findings on Technology, Organizational, and Policy Dimensions to Emergency Response and Related E-Governmental Services Δ	115
Horan, Thomas A.; Marich, Michael; Schooley, Ben	
Secure Interoperation for Effective Data Mining in Border Control and Homeland Security Applications Ω	124
Adam, Nabil R.; Atluri, Vijayalakshmi; Koslowski, Rey; Grossman, Robert; Janeja, Vandana P.; Warner, Janice	
Project Highlights: Multiple Agency and Jurisdiction Organized Response (M.A.J.O.R.) Disaster Research (NSF award # 428216) Ω	126
Batteau, Allen W.; Brandenburg, Dale; Brewster, Jon; Seeger, Matt; White, Suzanne	
SESSION 4C: E-CITIES	129
<i>Moderator: Genevieve Giuliano, University of Southern California, USA</i>	
The Fully Mobile City Government Project (MCity) Ω	131
Scholl, Hans (Jochen); Fidel, Raya; Mai, Jens-Erik	
UrbanSim: Interaction and Participation in Integrated Urban Land Use, Transportation, and Environmental Modeling Ω	133
Borning, Alan; Waddell, Paul	
Simulating Impact of Light Rail on Urban Growth in Phoenix: An Application of Urbanism Modeling Environment Δ	135
Guhathakurta, Subhrajit; Joshi, Himanshu; Konjevod, Goran; Crittenden, John; Li, Ke	
Intelligent Cities Ω	142
Curwell, Steve	
SESSION 5A: TECHNOLOGY TRANSFER	145
<i>Moderator: Rene Wagenaar, Delft University of Technology, The Netherlands</i>	
eGOVERNET - An European eGovernment Research Funding Agency Network Ω	147
Knudsen, Trond; Siösteen-Thiel, Madeleine	
Multidisciplinary E-Government Research and Education as a Catalyst for Effective Information Technology Transfer to Regional Governments Ω	149
Vélez-Rivera, Bienvenido; Fernández-Sein, Rafael; Rodríguez-Martínez, Manuel; Rivera-Vega, Pedro I.; Díaz, Walter; Nuñez-Molina, Mario	
Challenges in eGovernment Technology Transfer Ω	151
Chun, Soon Ae; Yesha, Yelena; Adam, Nabil; Atluri, Vijay	
COSPA (Consortium for studying, evaluating, and supporting the introduction of Open Source software and Open Data Standards in the Public Administration) Ω	153
Rossi, Bruno; Russo, Barbara; Succi Giancarlo	
SESSION 5B: E-RULEMAKING 1	155
<i>Moderator: Stuart Shulman, University of Pittsburgh, USA</i>	
Multidimensional Text Analysis for eRulemaking Δ	157
Kwon, Namhee; Shulman, Stuart W.; Hovy, Eduard	
Automatically Labeling Hierarchical Clusters Δ	167
Treeratpituk, Pucktada; Callan, Jamie	
Using Natural Language Processing to Improve eRulemaking Ω	177
Cardie, Claire; Farina, Cynthia; Bruce, Thomas; Wagner, Erica	

SESSION 5C: TRANSPARENCY AND E-GOVERNANCE	179
<i>Moderator: Luis F. Luna Reyes, Universidad de las Americas, Mexico</i>	
Transitioning from e-Government to e-Governance in the Knowledge Society: The Role of the Legal Framework for Enabling the Process in the European Union's Countries Δ	181
Paskaleva-Shapira, Krassimira	
A National Center for Digital Government Program on Networked Governance Ω	191
Fountain, Jane E.; Lazer, David	
Connecting to Congress Ω	193
Lazer, David; Esterling, Kevin; Neblo, Michael; Fountain, Jane; Mergel, Ines; Ziniel, Curt	
Digital Deliberation: Searching and Deciding About How to Vote Ω	195
Robertson, Scott P.	
SESSION 6A: STUDENT RESEARCH PRESENTATIONS	197
<i>Moderator: Sharon S. Dawes, University at Albany/SUNY, USA</i>	
The Social Relations of e-Government Diffusion in Developing Countries: The Case of Rwanda Δ	199
Mwangi, Wagaki	
e-Governance in Africa, from theory to action: a practical-oriented research and case studies on ICTs for Local Governance Δ	209
Misuraca, Gianluca Carlo	
Automated Classification of Congressional Legislation Δ	219
Purpura, Stephen; Hillard, Dustin	
SESSION 6B: E-RULEMAKING 2	227
<i>Moderator: José Luis Ambite, University of Southern California, USA</i>	
Locating Related Regulations Using a Comparative Analysis Approach Δ	229
Law, Kincho H.; Lau, Gloria T.; Wang, Haoyi	
Next Steps in Near-Duplicate Detection for eRulemaking Δ	239
Yang, Hui; Callan, Jamie; Shulman, Stuart	
Progress in Language Processing Technology for Electronic Rulemaking Ω	249
Shulman, Stuart; Callan, Jamie; Hovy, Eduard; Zavestoski, Stephen	
SESSION 6C: CRISIS MANAGEMENT 3	251
<i>Moderator: Jay Kesan, University of Illinois, USA</i>	
Secure-CITI Critical Information-Technology Infrastructure Ω	253
Mossé, Daniel; Comfort, Louise; Amer, Ahmed; Brustoloni, José C.; Chrysanthis, Panos K.; Hauskrecht, Milos; Labrinidis, Alexandros; Melhem, Rami; Pruh, Kirk	
E-Government and the Preparation of Citizens for Disasters Ω	255
Basolo, Victoria; Steinberg, Laura; Gant, Stephen	
SESSION 7A: PARTICIPATORY DESIGN AND MEDIATION	257
<i>Moderator: Marianne Lykke Nielsen, Royal School of Library and Information Science, Denmark</i>	
Developing a Youth-Services Information System for City and County Government: Experiments in User-Designer Collaboration Δ	259
Zappen, James P.; Adali, Sibel; Harrison, Teresa M.	
Policy Through Software Defaults Δ	265
Shah, Rajiv C.; Kesan, Jay P.	
Experimental Application of Process Technology to the Creation and Adoption of Online Dispute Resolution Ω	273
Sondheimer, Norman K.; Katsh, Ethan; Rainey, Daniel; Osterweil, Leon J.	

SESSION 7B: DIGITAL DOCUMENT PRESERVATION AND ARCHIVING	275
<i>Moderator: Laura Steinberg, George Washington University, USA</i>	
Building a state government digital preservation community: Lessons on interorganizational collaboration Δ	277
Kwon, Hyuckbin; Pardo, Theresa A.; Burke, G. Brian	
Robust Technologies for Automated Ingestion and Long-Term Preservation of Digital Information Ω	285
JaJa, Joseph	
Building a Demonstration Prototype for the Preservation of Large-Scale Multimedia Collections Ω	287
Rajasekar, Arcot; Berman, Francine; Burstan, Lynn; Kreisler, Harry; Schottlaender, Brian; Moore, Reagan; Marciano, Richard; Hou, Chien-Yi; Anderson, Steve; McEwen, Melissa; Bornheimer, Bee; DeClerck, Luc; Westbrook, Brad; Hutt, Arwen; Kozbial, Ardys; Fryman, Chris; Chu, Vivian	
DIGARCH Project Highlights: Multi-Institutional Testbed for Scalable Digital Archiving Ω	289
Miller, Stephen P.; Detrick, Robert S.; Helly, John	
SESSION 7C: SPATIO TEMPORAL AND GIS	291
<i>Moderator: Andrew Philpot, University of Southern California, USA</i>	
Voting Prediction Using New Spatiotemporal Interpolation Methods Δ	293
Gao, Jun; Revesz, Peter	
Scalable Data Collection and Retrieval Infrastructure for Digital Government Applications Ω	301
Samet, Hanan; Golubchik, Leana	
Automatic Alignment of Vector Data and Orthoimagery for The National Map Ω	303
Knoblock, Craig A.; Shahabi, Cyrus; Chen, Ching-Chien; Usery, E. Lynn	
National Large-Scale Urban True Orthophoto Mapping and Its Standard Initiative Ω	305
Zhou, Guoqing; Xie, Wenhan; Benjamin, Susan; Fegeas, Robin G.; Simmers, John; Cluff, Hap; Lei, Y.; Foust, Jeanne	
SESSION 8A: PROCESS AND WORKFLOW	307
<i>Moderator: Lois Delcambre, Portland State University, USA</i>	
Lynx: An Open Architecture for Catalyzing the Deployment of Interactive Digital Government Workflow-Based Systems Δ	309
Vélez, Iván P.; Vélez, Bienvenido	
Argos: Dynamic Composition of Web Services for Goods Movement Analysis and Planning (Abstract) Ω	319
Ambite, José Luis; Giuliano, Genevieve; Gordon, Peter; Pan, Qisheng; Jinwala, Mountu; Kapoor, Dipsy; Wang, LanLan	
Data Processing Workflows in the Social Sciences: Representation and Automatic Generation ◊	321
Ambite, José Luis; Kapoor, Dipsy; Jinwala, Mountu	
SESSION 8B: FEDERAL AGENCIES AND THE WEB	321
<i>Moderator: J. Ramon Gil-Garcia, University at Albany/SUNY, USA</i>	
Federal Agencies and the Evolution of Web Governance Δ	323
Mahler, Julianne; Regan, Priscilla M.	
Data Confidentiality, Data Quality and Data Integration for Federal Databases Ω	325
Karr, Alan F.	
Integrating Data and Interfaces to Enhance Understanding of Government Statistics: Toward the National Statistical Knowledge Network Project Briefing Ω	334
Marchionini, Gary; Haas, Stephanie; Plaisant, Catherine; Shneiderman, Ben	

SESSION 8C: INTERNATIONAL DIGITAL GOVERNMENT PROJECTS	337
<i>Moderator: Jane Fountain, University of Massachusetts, Amherst, USA</i>	
Accelerated Indexing in a Domain-Specific Digital Library Ω	339
Delcambre, Lois; Price, Susan; Nielsen, Marianne Lykke; Tolle, Timothy; Luk, Vibeke; Weaver, Mathew	
LOG-IN Africa: Local Governance and ICTs Research Network for Africa Ω	341
Misuraca, Gianluca	
Building efficiency through ICT utilization in the Government of Japan Ω	342
Okumura, Hirokazu	
SYSTEM DEMONSTRATIONS	345
DURIAN: A Demo for Near-Duplicate Detection	347
Callan, Jamie; Shulman, Stuart; Yang, Hui	
webADIS: A Flexible web-based Environment for the Automated Dental Identification System Ω	348
Chekuri, Satyashrinivas; Nassar, Diaa Eldin; Abaza, Ayman; Haj Said, Eyad; Bahu, Ali; Qurashi, Uthman; Fahmy, Gamal; Ammar, Hany	
Living on the Edge with the Oregon Coastal Atlas	350
Klarin, Paul; Haddad, Tanya; Cone, Joseph; Wright, Dawn J.	
Integrating Information Technology and Social Science Research for Effective Government: MOST Policy Research Tool	352
Maugis, Vincent	
Integration of GIS and Educational Achievement Data for Education Policy Analysis and Decision-making	354
Mulvenon, Sean W.; Wang, Kening; McKenzie, Sarah; Airola, Denise; Anderson, Travis	
A Process-Driven Tool to Support Online Dispute Resolution	356
Sondheimer, Norman K.; Osterweil, Leon J.; Katsh, Ethan; Clarke, Lori; Gaithenby, Alan; Wing, Leah; Rainey, Daniel; Marzilli, Matthew; Wise, Alexander; Gyllstrom, Daniel	
Supporting Humanitarian Relief Logistics Operations through Online Geocollaborative Knowledge Management	358
Tomaszewski, Brian M.; MacEachren, Alan M.; Pezanowski, Scott; Liu, Xiaoyan; Turton, Ian	
Opus (the Open Platform for Urban Simulation) and UrbanSim 4	360
Waddell, Paul; Borning, Alan; Ševcíková, Hana; Socha, David	
POSTERS	363
Constituent-centric Municipal Government Coalition Portal ♦	365
Adam, Nabil R.; Atluri, Vijay; Chun, Soon Ae; Artigas, Francisco; Bora, Irfan; Ceberio, Bob	
Semantics-based Threat Structure Mining ♦	367
Adam, N.; Atluri, V.; Janeja, P.; Paliwal, A.; Youssef, M.; Chun, S.; Cooper, J.; Paczkowski, J; Bornhoevd, C.; Nassi, I.; Schaper, J.	
Automated Dental Identification System (ADIS) Ω	369
Ammar, Hany; Howell, Robert; Abdel-Mottaleb, Mohamed; Jain, Anil	
An Empirical Study on E-Government Readiness: The Roles of Institutional Efficiency and Interpersonal Trust ♦	371
Bagchi, Kallol; Galup, Stuart; Cerveny, Robert	
The BioPortal Project: A National Center of Excellence for Infectious Disease Informatics Ω	373
Chen, Hsinchun; Zeng, Daniel; Tseng, Chunju; Larson, Cathy	
Understanding the Adoption and Diffusion of Innovative Information Technology Curricula: A Case Application to Master of Public Administration Programs ♦	375
Chiu, Shu-Chuan	

Quality Evaluation of e-Government Digital Services ◊	377
Corradini, Flavio; De Angelis, Francesco; Polzonetti, Alberto; Re, Barbara	
The Role of Public Return on Investment Assessment in Government IT Projects ◊	379
Cresswell, Anthony M.	
Eco-Informatics and Natural Resource Management Ω	381
Cushing, Judith Bayard; Wilson, Tyrone	
Overview: Building a Sustainable International Digital Government Research Community Ω	383
Dawes, Sharon S; Gregg, Valerie	
Technology Adoption and Institutional Change in the United States Senate: An Analysis of Web Site Content ◊	385
Esterling, Kevin; Lazer, David; Neblo, Michael	
eGovernment for Business across the Atlantic: from Cases to Models ◊	387
Fariselli, Patrizia; Culver-Hopper, Julia; Bojic, Olana	
SGER: Project Summary - CAPWIN Ω	390
Gaynor, Mark	
Improving the Workflow while Reducing the Costs: Using XML for Web Site Content Management in Government Agencies ◊	392
Gil-Garcia, J. Ramon; Canestraro, Donna; Costello, Jim; Baker, Andrea; Werthmuller, Derek	
Enacting Inter-Organizational E-Government in the Mexican Federal Government ◊	394
Gil-Garcia, J. Ramon; Luna-Reyes, Luis Felipe	
Estimating Freight Flows for Metropolitan Highway Networks Using Secondary Data Sources ◊	396
Giuliano, Genevieve; Gordon, Peter; Pan, Qisheng; Park, Jiyoung; Wang, LanLan	
TIME-CRITICAL INFORMATION SERVICES: Update on Exploratory Analysis of Emergency Response and Related E-Governmental Services Ω	398
Horan, Thomas A.; Marich, Michael; Schooley, Ben	
Entity Consolidation and Alignment in Semi-Structured Data Sources Ω	400
Hovy, Eduard; Philpot, Andrew; Pantel, Patrick	
UNeGov.net – Community of Practice for Electronic Governance ◊	402
Janowski, Tomasz; Estevez, Elsa; Khan, Irshad; Ojo, Adegboyega	
Unraveling Shared Services using Simulation ◊	404
Janssen, Marijn; Wagenaar, René W.	
Modeling and Forecasting of e-Vilnius Development ◊	406
Kaklauskas, Arturas	
Modeling Online Participation in Local Governance Ω	408
Kavanaugh, Andrea; Pérez-Quiñones, Manuel; Isenhour, Philip; Dunlap, Daniel	
Target Vehicle Identification for Border Safety with Modified Mutual Information ◊	410
Kaza, Siddharth; Wang, Yuan; Chen, Hsinchun	
Friends, Foes, and Fringe: Norms and Structure in Political Discussion Networks ◊	412
Kelly, John W.; Fisher, Danyel; Smith, Marc	
Electronic Government Capacity and Federal Program Performance: An Analysis of OMB's PART Scores and Executive Branch Management Scorecard ◊	418
Kim, Hyun Joon; Kim, Soonhee	
Research and Development for Innovative Government – A National Agenda for Renewal Ω	420
Knudsen, Trond	
A Distributed Information Management Framework (REGNET) for Environmental Laws and Regulations Ω	423
Law, Kincho H.	
Citizen Centric Analysis of Anti/Counter-Terrorism e-Government Services Ω	425
Lee, JinKyu; Rao, H. R.	

Periodic Association Mining in a Geospatial Decision Support System ◊	427
Li, Dan; Deogun, Jitender S.	
Digitalization of Coastal Management and Decision Making Supported by Multi-Dimensional Geospatial Information and Analysis Ω	429
Li, Ron; Bedford, Keith; Shum, C.K.; Niu, Xutong; Zhou, Feng; Velissariou, Vasilia; Ramirez, J. Raul; Zhang, Aidong	
Locating Online Government Information: A Comparison of FirstGov, Google, and Yahoo ◊	431
Meho, Lokman I.; Yang, Kiduk	
Interactive Design Best Practices for the Public Sector ◊	433
Miller, Eric	
Multi-Institution Testbed for Scalable Digital Archiving Ω	435
Miller, Stephen P.; Detrick, Robert; Helly, John	
Repository Replication Using SMTP and NNTP Ω	436
Nelson, Michael L.; Smith, Joan A.; Klein, Martin	
Matching and Integration Across Heterogeneous Data Sources ◊	438
Pantel, Patrick; Philpot, Andrew; Hovy, Eduard	
The Impacts of Digital Government on Civic Engagement: A Typology of Information Technology Use ◊	440
Park, Hun Myoung	
Building Regulatory Compliant Storage Systems Ω	442
Peterson, Zachary N. J.; Burns, Randal	
Regionalizing Integrated Watershed Management: A Strategic Vision ◊	444
Pezzoli, Keith; Marciano, Richard; Robertus, John	
Distributed Higher-Order Text Mining: Theory and Practice Ω	446
Pottenger, William M.; Li, Shenzhi; Janneck, Christopher D.	
Scalable and Secure Data Collection: Incentives Considerations ◊	448
Raveendran, Ranjit; Cheng, William C.; Golubchik, Leana	
Elements of Social Science Engagement in Information Infrastructure Design ◊	450
Ribes, David; Baker, Karen S.	
Effective Citizen Relationship Management: Hurricane Wilma and Miami-Dade County 311 ◊	452
Schellong, Alexander; Langenberg, Thomas	
What can e-Commerce and e-Government Learn from Each Other? ◊	454
Scholl, Hans J (Jochen)	
Virtualization Technologies in Transnational DG ◊	456
Tsugawa, Maurício; Matsunaga, Andrea; Fortes, José A. B.	
An Electronic Social Network to Market Topics of Public Interest: Net@INA ◊	458
Valadares Tavares, L.; Silva, Paulo	
A Performance Ratings Framework for the Evaluation of Electronic Voting Systems Ω	460
Vora, Poorvi L.; Simha, Rahul; Stanton, Jonathan	
A Probabilistic Model for Approximate Identity Matching ◊	462
Wang, G. Alan; Chen, Hsinchun; Atabakhsh, Homa	
Semantic Web Technologies to Automate Searching for Geospatial Data ◊	464
Wiegand, Nancy	
Design Principles for Public Safety Response Mobilization ◊	466
Williams, Christine B.; Fedorowicz, Jane; Markus, M. Lynne; Sawyer, Steve; Tyworth, Michael	
University Information System RUSSIA: data, knowledge products and services for social research	468
Yudina, Tatyana	
Connected Kids: Designing a Youth-Services Information System for Local Government Ω	470
Zappen, James P.; Adali, Sibel; Harrison, Teresa M.	

Accuracy Improvement of Urban True Orthoimage Generation Using 3D R-tree-based Urban Model ♦ Zhou, Guoqing; Xie, Wenhan	472
BIRDS-OF-A-FEATHER	479
Interdisciplinary Analysis of Digital Government Work	481
Scholl, Hans J (Jochen); Mai, Jens-Erik; Fidel, Raya	
E-Government Measurement and Evaluation	484
Luna-Reyes, Luis Felipe; Gil-Garcia, J. Ramon	
XML for Web Site Management in Government: State of the Art and Future Research	485
Gil-Garcia, J. Ramon; Canestraro, Donna; Costello, Jim; Baker, Andrea; Werthmuller, Derek	
e-Governance as a Global Knowledge-Enabling Platform	487
Finger, Matthias; Misuraca, Gianluca; Rossel, Pierre	
Using System Dynamics for Theory Building in Digital Government Research: Exploring the Dynamics of Digital Government Evolution	488
Martinez-Moyano, Ignacio J.	
Citizen Relationship Management: Understanding, Challenges and Impact	490
Schellong, Alexander	
FIELD TRIPS	491
The Synthesis Center	493
U.S. Border Patrol Command and Control Intelligence Coordination Center	495
ABOUT THE DIGITAL GOVERNMENT RESEARCH CENTER	497

Conference Program at a Glance for dg.o 2006

SUNDAY	8:00am-9:00am	Registration
	9:00am-5:00pm	Workshops and Tutorials
	6:00pm-9:00pm	Pls Reception
	6:00pm-9:00pm	Student Reception

	MONDAY	TUESDAY	WEDNESDAY
7:00	Registration and continental breakfast		
7:30		Registration and continental breakfast	Registration and continental breakfast
8:00	Plenary Keynote Address: Dr. Fran Berman, San Diego Super Computer Center		
8:30		Plenary Keynote Address: Dr. John Phillips, CIA	Research Presentations (7A): Participatory Design and Mediation
9:00			Research Presentations (7B): Digital Document Preservation and Archiving
9:30	Break 9:30 - 10:00 am	Break 9:30 - 9:45 am	Research Presentations (7C): Spatio Temporal and GIS
9:45		Research Presentations (4A): Crisis Management 2 Panel (4B): Sustaining an International Digital Government Research Community Panel (4C): e-Cities	
10:00	Research Presentations (1A): Data Mining		Break 10:00 - 10:30 am
10:30	Research Presentations (1B): Voting		Research Presentations (8A): Process and Workflow
11:00	Panel (1C): Cyberinfrastructure for Public Health: Digital Government Research Collaborations		Research Presentations (8B): Federal Agencies and the Web Research Presentations (8C): International Digital Government Projects
11:30	Break 11:30 - 11:45 am	Lunch (On Your Own) 11:30 am - 12:45 pm	
11:45			NSF DG PI's Luncheon (Others On Their Own) 12:00 - 1:00 pm
12:00			
12:45	Birds-of-a-Feather (BOF) Sessions		
1:00		Plenary Keynote Address: Kimberly Nelson, Microsoft	
1:15	Plenary Panel		
1:30	National Science Foundation Program Managers: Perspectives on Sustaining Digital Government Research		
1:45		Break 1:45 - 2:00 pm	
2:00	Break 2:00 - 2:15 pm	Research Presentations (5A): Technology Transfer	
2:15	Research Presentations (2A): Integrated Justice	Research Presentations (5B): e-Rulemaking 1 (Text Analysis)	
2:45			
3:00	Research Presentations (2B):		Afternoon Activities from 1:30 - 5:30 pm listed below: International Meeting Roadmap 2020 Field Trips: The Synthesis Center

3:00	<u>Citizen Participation 1</u> <u>Panel (2C): Ecosystem Informatics for Decision-Makers</u>	<u>Research Presentations (5C): Transparency and e-Governance</u> Break 3:30 - 3:45 pm	US Border Control Center Afternoon Workshops and Tutorials
3:45	Research Presentations (3A): <u>Crisis Management 1</u>	Research Presentations (6A): <u>Student Research Presentations</u>	
4:00	Research Presentations (3B): <u>Citizen Participation 2</u>	Research Presentations (6B): <u>e-Rulemaking 2</u>	
4:30	Research Presentations (3C): <u>Emergent Design</u>	Research Presentations (6C): <u>Crisis Management 3</u>	
4:45	Break 4:45 - 5:00 pm		
5:00	<u>Plenary: Digital Government Society</u>	Break 5:00 - 6:00 pm	
5:45			
6:00		Poster/Demo with Reception (Dinner Provided)	
7:00			
7:30		Break 7:30 - 8:00 pm	
8:00		Poster/Demo with Reception (Dinner Provided)	

Author Index

- Abaza, Ayman, 348
Abdel-Mottaleb, Mohamed, 369
Adali, Sibel, 259, 470
Adam, Nabil R., 124, 151, 365, 367
Airola, Denise, 354
Ambite, José Luis, 319, 321
Amer, Ahmed, 253
Ammar, Hany, 348, 369
Anderson, Steve, 287
Anderson, Travis, 354
Artigas, Francisco, 365
Atabakhsh, Homa, 49, 462
Atluri, Vijayalakshmi, 124, 151, 365, 367
Bagchi, Kallol, 371
Bahu, Ali, 348
Baker, Andrea, 392, 485
Baker, Karen S, 450
Baru, Chaitan, 21
Basolo, Victoria, 255
Batteau, Allen W., 126
Bedford, Keith, 429
Benjamin, Susan, 305
Berman, Francine, 3, 287
Biasiotti, Maria Angela, 62
Bojic, Olana, 387
Bora, Irfan, 365
Bornheimer, Bee, 287
Bornhoevd, C., 367
Borning, Alan, 133, 360
Brandenburg, Dale, 126
Brandt, Lawrence E., *vii*, 11
Brazile, Robert, 45
Breuer, Fabian, 40
Brewster, Jon, 126
Brooks, Terry, 67
Bruce, Thomas, 177
Brustoloni, José C., 253
Burke, G. Brian, 277
Burns, Randal, 442
Burstan, Lynn, 287
Cai, Guoray, 71
Callan, Jamie, 167, 239, 249, 347
Canestraro, Donna, 392, 485
Cappella, Joseph N., 89
Cardie, Claire, 177
Caroll, John M., 79
Ceberio, Bob, 365
Cerveny, Robert, 371
Chekuri, Satyashrinivas, 348
Chen, Ching-Chien, 303
Chen, Hsinchun, 49, 373, 410, 462
Cheng, William C., 448
Chiu, Shu-Chuan, 375
Chrysanthis, Panos K., 253
Chu, Vivian, 287
Chun, Soon Ae, 151, 365, 367
Clarke, Lori, 356
Cluff, Hap, 305
Comfort, Louise, 253
Cone, Joseph, 350
Contractor, Noshir, 9
Cooper, J., 367
Corradini, Flavio, 377
Costello, Jim, 392, 485
Creamer, German, 23
Cresswell, Anthony M., 379
Crittenden, John, 135
Culver-Hopper, Julia, 387
Curwell, Steve, 142
Cushing, Judith Bayard, 12, 381
Dawes, Sharon S., 14, 383
De Angelis, Francesco, 377
DeClerck, Luc, 287
Delcambre, Lois, 25, 339
Deogun, Jitender S., 427
Detrick, Robert S., 289, 435
Díaz, Walter, 149
Dunlap, Daniel, 408
Esterling, Kevin, 193, 385
Estevez, Elsa, 402
Fahmy, Gamal, 348
Farina, Cynthia, 177
Fariselli, Patrizia, 387
Fedorowicz, Jane, 466
Fegeas, Robin G., 305
Fernández-Sein, Rafael, 149
Fidel, Raya, 131, 481
Finger, Matthias, 487
Fisher, Danyel, 412
Fortes, José A. B., 456
Fountain, Jane E., 191, 193
Fountain, Tony, 21
Foust, Jeanne, 305
Fryman, Chris, 287
Fuhrmann, Sven, 71
Gaithenby, Alan, 356
Galup, Stuart, 371
Gant, Stephen, 255
Gao, Jun, 293
Gaynor, Mark, 390

- Gil-Garcia, J. Ramon, 392, 394, 484, 485
Giuliano, Genevieve, 319, 396
Golubchik, Leana, 301, 448
Gordon, Peter, 319, 396
Govindaraju, Venu, 73
Gregg, Valerie J., 14, 383
Grossman, Robert, 124
Guhathakurta, Subhrajit, 135
Gyllstrom, Daniel, 356
Haas, Stephanie, 334
Haddad, Tanya, 350
Haj Said, Eyad, 348
Harrison, Teresa M., 259, 470
Hauskrecht, Milos, 253
Helly, John, 289, 435
Herrnson, Paul S., 38
Hershkop, Shlomo, 23
Hesse, Bradford W., 9
Hillard, Dustin, 219
Hodgkiss, William, 21
Horan, Thomas A., 115, 398
Hou, Chien-Yi, 287
Hovy, Eduard, 157, 249, 400, 438
Howell, Robert, 369
Hutt, Arwen, 287
Isenhour, Philip, 79, 408
Jackson, Steven, 95
Jain, Anil, 369
JaJa, Joseph, 285
Janeja, Vandana P., 124, 367
Jankowski, Piotr, 67
Janneck, Christopher D., 446
Janowski, Tomasz, 402
Janssen, Marijn, 404
Jasso, Hector, 21
Jinwala, Mountu, 319, 321
Joshi, Himanshu, 135
Joshi, Shailesh, 49
Kaklauskas, Arturas, 406
Kapoor, Dipsy, 319, 321
Karr, Alan F., 332
Katsh, Ethan, 273, 356
Kavanaugh, Andrea, 79, 408
Kaza, Siddharth, 49, 410
Kelly, John W., 91, 412
Kesan, Jay P., 265
Khan, Irshad, 402
Kim, Hyun Joon, 418
Kim, Soonhee, 418
Klarin, Paul, 350
Klein, Martin, 436
Knoblock, Craig A., 303
Knudsen, Trond, 147, 420
Konjevod, Goran, 135
Koslowski, Rey, 124
Kozbial, Ardys, 287
Kreisler, Harry, 287
Kwon, Hyuckbin, 277
Kwon, Namhee, 157
Labrinidis, Alexandros, 253
Langenberg, Thomas, 452
Larson, Cathy, 373
Lau, Gloria T., 229
Law, Kincho H., 229, 423
Lazer, David, 191, 193, 385
Lee, JinKyu, 425
Lei, Y., 305
Li, Dan, 427
Li, Ke, 135
Li, Ron, 429
Li, Shenzhi, 446
Liu, Xiaoyan, 358
Luk, Vibeke, 25, 339
Luna-Reyes, Luis Felipe, 394, 484
MacEachren, Alan M., 71, 358
Mahler, Julianne, 325
Mai, Jens-Erik, 131, 481
Marchionini, Gary, 334
Marciano, Richard, 287, 444
Marich, Michael, 115, 398
Markus, M. Lynne, 466
Martin, Fred, 12
Martinez-Moyano, Ignacio J., 488
Marzilli, Matthew, 356
Matsunaga, Andrea, 456
Maugis, Vincent, 352
McEwen, Mellisa, 287
McKenzie, Sarah, 354
McNeese, Michael, 71
Meho, Lokman I., 431
Melhem, Rami 253
Mergel, Ines, 193
Miller, Eric, 433
Miller, Stephen P., 289, 435
Misuraca, Gianluca Carlo, 209, 341, 487
Moore, Reagan, 287
Mossé, Daniel, 253
Muhlberger, Peter, 53
Mulvenon, Sean W., 354
Mwangi, Wagaki, 199
Nannucci, Roberta, 62
Nassar, Diaa Eldin, 348
Nassi, I., 367
Neblo, Michael, 193, 385

- Nelson, Kimberly, 5
Nelson, Michael L., 436
Nielsen, Marianne Lykke, 25, 339
Niemi, Richard G., 38
Niu, Xutong, 429
Nuñez-Molina, Mario, 149
Nyerges, Tim, 67
Ojo, Adegboyega, 402
Okumura, Hirokazu, 342
Osterweil, Leon J., 273, 356
Paczkowski, J., 367
Paliwal, A., 367
Pan, Qisheng, 319, 396
Pantel, Patrick, 400, 438
Pardo, Theresa A., 12, 277
Park, Hun Myoung, 440
Park, Jiyoung, 396
Paskaleva-Shapira, Krassimira, 181
Patil, G.P., 75
Pérez-Quiñones, Manuel, 79, 408
Petersen, Tim, 49
Peterson, Zachary N. J., 442
Pezanowski, Scott, 358
Pezzoli, Keith, 444
Philpot, Andrew, 400, 438
Plaisant, Catherine, 334
Polzonetti, Alberto, 377
Pottenger, William M., 446
Price, Susan, 25, 339
Price, Vincent, 89
Pruhs, Kirk, 253
Purpura, Stephen, 219
Qurashi, Uthman, 348
Rainey, Daniel, 273, 356
Rajasekar, Arcot, 287
Ramirez, J. Raul, 429
Rao, H. R., 425
Raveendran, Ranjit, 448
Re, Barbara, 377
Regan, Priscilla M., 325
Reich, Don, 21
Revesz, Peter, 293
Ribes, David, 450
Rivera-Vega, Pedro I., 149
Robertson, Scott P., 195
Robertus, John, 444
Rodríguez-Martínez, Manuel, 149
Rossel, Pierre, 487
Rossi, Bruno, 153
Russ, Barbara, 153
Rutherford, G. Scott, 67
Samet, Hanan, 301
Sawyer, Steve, 47, 105, 466
Schaper, J., 367
Schellong, Alexander, 452, 490
Schmitz, Joseph, 79
Schnase, John, 12
Scholl, Hans J (Jochen), 131, 454, 481
Schooley, Ben, 115, 398
Schottlaender, Brian, 287
Seeger, Matt, 126
Ševciková, Hana, 360
Shah, Rajiv C., 265
Shahabi, Cyrus, 303
Sharma, Rajeev, 71
Shneiderman, Ben, 334
Shulman, Stuart W., 157, 239, 249, 347
Shum, C.K., 429
Silva, Paulo, 458
Simha, Rahul, 460
Simmers, John, 305
Siösteen-Thiel, Madeleine, 147
Smith, Joan A., 436
Smith, Marc, 412
Socha, David, 360
Sondheimer, Norman K., 273, 356
Spengler, Sylvia, 11, 12
Stanton, Jonathan, 460
Stark, David, 91
Steinberg, Laura, 255
Stolfo, Salvatore J., 23
Succi, Giancarlo, 153
Sugarbaker, Larry, 12
Swigger, Kathleen, 45
Tolle, Timothy, 25, 339
Tomaszewski, Brian M., 358
Traugott, Michael W., 38
Treichsel, Alexander H., 40
Treeratpituk, Pucktada, 167
Tseng, Chunju, 373
Tseng, Lu Chunju, 49
Tsugawa, Maurício, 456
Turton, Ian, 358
Tyworth, Michael, 47, 105, 466
Usery, E. Lynn, 303
Valadares Tavares, L., 458
Vélez, Bienvenido, 309
Vélez, Iván P., 309
Vélez-Rivera, Bienvenido, 149
Velissariou, Vasilia, 429
Violette, Chuck, 49
Vora, Poorvi L., 460
Waddell, Paul, 133, 360
Wagenaar, René W., 404

- Wagner, Erica, 177
Wang, Alan G., 49
Wang, G. Alan, 462
Wang, Haoyi, 229
Wang, Kening, 354
Wang, LanLan, 319, 396
Wang, Yuan, 49, 410
Warner, Janice, 124
Warner, Kurt, 21
Weaver, Mathew, 25, 339
Werthmuller, Derek, 392, 485
Westbrook, Brad, 287
White, Suzanne, 126
Wiegand, Nancy, 464
Williams, Christine B., 466
Wilson, Tyrone, 12, 381
Wing, Leah, 356
Wise, Alexander, 356
- Wolber, David, 29
Wright, Dawn J., 350
Xie, Wenhan, 305, 472
Yang, Hui, 239, 347
Yang, Kiduk, 431
Yesha, Yelena, 151
Young, Rhonda, 67
Youssef, M., 367
Yudina, Tatyana, 468
Zappen, James P., 259, 470
Zavestoski, Stephen, 249
Zeng, Daniel, 373
Zhang, Aidong, 429
Zhou, Feng, 429
Zhou, Guoqing, 305, 472
Zin, Than Than, 79
Ziniel, Curt, 193

INVITED TALKS

Titles and Authors

One Hundred Years of Data
Fran Berman, University of California San Diego, USA

Academic and Business Partnerships to Enhance Digital Government Research
Kimberly Nelson, Microsoft Corporation, USA

One Hundred Years of Data

Dr. Fran Berman
Director
San Diego Supercomputer Center
Professor and High Performance Computing Endowed Chair
UC San Diego

ABSTRACT:

The 20th century brought about an “information revolution” which has forever altered the way we work, communicate, and live. In the 21st century, it is hard to imagine working without an increasingly broad array of enabling technologies and the data they provide. Much of this data will form the foundation for new discovery, advances, and policy over the next 100 years and beyond.

The care and management of today's tidal wave of data has become an increasingly important focus for technology development. Collecting, providing, and preserving data responsibly presents both an opportunity and a challenge. Whereas books can be preserved for years and even centuries, the preservation of digital data is dependent on the technologies on which it is stored. In the next 100 years, storage technologies will advance tens of generations, and the digital collections preserved on up - to - date storage technologies will need to transition through each new generation, and many times over.

Without a planned approach to preservation, valuable data will be damaged or lost. The stakes are high – some data collections such as the Shoah Collection of Holocaust survivor testimony are irreplaceable, and some data collections such as the longitudinal Panel Study of Income Dynamics used by social scientists, and the Protein Data Bank used by biologists, are fundamental research tools. The challenges of responsible data preservation are great. Key questions that must be addressed in the preservation of long - lived digital data include:

1) What should we save?

We can't save everything, and even if we could, it would be exceedingly difficult to find useful information within the mass of data. Some data collections will need to be marked for preservation from the outset, and some collections will need to be “rescued”.

2) Who is responsible?

Digital collections are of interest to many constituents - - data generators, users, stewards, etc. Who is responsible for preserving

the digital data over the long - term? Who will pay for upkeep, technology transition, and the development of tools and interfaces to make the data accessible?

3) How do we keep data safe?

Digital media is more fragile than paper. Software bugs, power outages, hackers, and other problems threaten the reliability of digital collections. The risks can be mitigated when multiple copies of the data collection are generated and updated consistently.

4) How should we save it?

Communities vary widely in their usage patterns, formats, standards, and policies with respect to digital data. The cyberinfrastructure in

which digital repositories are embedded must provide reliable, safe and usable access. Data collections must be available for analysis, modeling, public access, dissemination, and other types of usage.

These questions and many other challenges must be addressed for responsible digital preservation. In this talk we focus on the development and deployment of Cyberinfrastructure for data management and preservation, and the challenges of developing a framework for data management and preservation over the foreseeable future.

Academic and Business Partnerships to Enhance Digital Government Research

Kimberly T. Nelson
Executive Director, EGovernment
U.S. Public Sector
Microsoft Corporation
5335 Wisconsin Ave, N.W. Suite 600
Washington, DC 20015
202-895-7468 office

ABSTRACT:

Kimberly Nelson is Executive Director for eGovernment at Microsoft Corporation where she is helping develop Microsoft's e-government strategy. She will talk about how Microsoft is working with partners to establish long-term strategies for more efficient and cost-effective online services and government solutions.

Ms. Nelson will discuss some of the trends that are impacting government policy makers today – increasingly mobile work forces, heightened demands for collaboration among government agencies, and the changing demographics of the people we serve and those who serve them, and the rise of “Government by Networks.” .

She will talk about policies and tools that the workforce of the future will need – to meet changes that are expected within organizations for both internal and external service delivery and effective decision making. Included in this discussion is the evolution from traditional eGovernment services to a true Digital Society.

Finally, Ms Nelson will discuss the interdisciplinary nature of the today's workforce -- governments, businesses and academia working together to solve government's toughest problems with an emphasis on academic and business partnerships to enhance digital government research.

PANELS

Topics and Moderators

SESSION 1C *Cyberstructure for Public Health: Digital Government Research Collaborators*
Moderator: Noshir Contractor, University of Urbana-Champaign, USA

PLENARY *National Science Foundation Program Managers: Perspectives on Sustaining
Digital Government Research*
Moderator: Valerie Gregg, University of Southern California, USA

SESSION 2C *Ecosystem Informatics for Decision-Makers*
Moderator: Tyrone Wilson, USGS National Biological Information Infrastructure, USA

SESSION 4B *Sustaining an International Digital Government Research Community*
Moderator: Sharon S. Dawes, University at Albany/SUNY, USA

Cyberinfrastructure for Public Health

Noshir Contractor, Ph.D.
Professor, Speech Communication &
Psychology

Director, Science of Networks in
Communities (SONIC) Group, NCSA
Univ of Illinois at Urbana-Champaign
702 South Wright Street
Urbana, IL 61801
01-217-333-7780
nosh@uiuc.edu

Bradford W. Hesse, Ph.D.
Chief, Health Communication and
Informatics Research Branch
National Cancer Institute, National
Institutes of Health
6130 Executive Blvd., MSC 7365
Bethesda, MD 20892-7365
01-301-594-9904
hesseb@mail.nih.gov

ABSTRACT

General Terms

Grid Computing, Health Informatics

Keywords

Interoperable Systems, Knowledge Management, Public Health, Networks, Surveillance, Behavioral Research

Background

Great strides were made during the 20th Century to improve the diagnosis and treatment of many common diseases, but those strides are not enough. Millions die each year from diseases that are chronic and complex. To meet the challenge of combating these complex diseases, biomedical research in the 21st Century must take advantage of advanced discovery in information technology to create interventions that are predictive, personalized, and preemptive. [1] At the patient level, 21st Century medicine must use the precision of evidence-based diagnostic systems to deliver highly tailored treatment regimens in precise, effective ways. At the population level, 21st Century disease control must use the power of connective surveillance infrastructures to identify targets of opportunity early, and to apply current knowledge for intervention in rapidly diffusing ways. To achieve these goals, population health in the 21st century must rely on the connective power of powerful health informatics infrastructures, or cyberinfrastructures in health.[2]

One of the most compelling cases for applying the power of cyberinfrastructure to address the challenges of 21st Century medicine can be found in the area of cancer control and prevention. What we now know as cancer is actually a complex family of diseases with similar but distinct etiologies and pathways for prevention.[3] Fortunately, the biggest threats from

cancer – lung cancer, colon cancer, breast cancer, cervical cancer, prostate cancer – can be addressed now, using current scientific knowledge.[4] It has been estimated that by simply delivering the benefits of current scientific knowledge to the population at large, the number of deaths due to cancer could be reduced by more than 50%. With cancer eclipsing heart disease as the leading cause of death among Americans under the age of 85 in 2004, the public health imperative for harnessing the power of cyberinfrastructure to prevent and control cancer has never been more important.[5]

The goal of the symposium will be to explore the ways in which new applications in cyberinfrastructure can be used to harness the power of discovery in cancer control and prevention. The symposium will bring together pioneers in the field who are creating applications to expand the capacity of the National Cancer Institute's Bioinformatics Grid into areas of direct relevance to population health. The symposium builds on a similar panel convened during the 2005 meeting of digital government grantees, but expands the analysis by offering concrete examples of how applications in cyberinfrastructure technology can be leveraged to elevate research in public health.

Speakers

Peter Schad, Booz Allen Hamilton (on contract to National Cancer Institute): The **cancer Biomedical Informatics Grid**, or **caBIG™**

Stephen Marcus, National Cancer Institute: Developing cyberinfrastructure for public health surveillance systems (HINTS & ToBIG)

Shu-Hong Zhu, Associate Adjunct Professor, Family & Preventive Medicine, Cancer Prevention & Control Program,

Moores cancer Center, University of California at San Diego: A Database Infrastructure to Reach and Assist Underserved Smokers through Quitlines

Noshir Contractor, Professor, University of Illinois at Urbana-Champaign and Director, Science of Networks in Communities (SONIC), National Center for Supercomputing Applications: Using Cyberinfrastructure to Enable Networks in Public Health

Patty Mabry, Health Scientist Administrator/Behavioral Scientist, Office of Behavioral and Social Sciences Research, Office of the Director, National Institutes of Health: NIH Plans for Further Development of Cyberinfrastructure to Enable Behavioral and Social Sciences Research.

Symposium Summary

Peter Schad will begin the symposium with a description of the vision and architecture underlying the National Cancer Institute's **cancer Biomedical Informatics Grid**, or *caBIG™*. The *caBIG* architecture is designed to bring the terabytes and petabytes of data being produced in cancer research onto a common, interoperable platform to enhance discovery, development, and delivery in the national program's war on cancer.

Stephen Marcus will describe how the NCI is exploring the use of cyberinfrastructure to connect data from relevant cancer surveillance systems into a seamless thread of support for public health researchers, policy makers, and public health administrators. Dr. Marcus will explain how the NCI has been making data from its nationally representative Health Information National Trends Survey (HINTS) available to communication researchers through an online collaboratory, and he will describe plans to link national data systems on tobacco use to improve the effectiveness and reach of public health surveillance.

Shu-Hong Zhu will offer an example of how large scale databases and metadata repositories can be used to enable research on data collected through a national consortium of publicly funded "Quitlines" for smokers. Because smoking-related cancers kill more Americans than any other cancer, increasing the reach and effectiveness of cessation resources must remain a public health priority. Creating a cyberinfrastructure to connect here-to-fore disparate data systems should elevate the precision of scientific

analysis in the area of *Quitline* research from a local to a national level.

Noshir Contractor, a current digital government grantee, will present a blueprint for connecting public health researchers through a distributed network of people, data, and resources in the area of public health surveillance and evaluation. Building on his work in social network analysis, Dr. Contractor will illustrate how members of a geographically distributed community of population scientists can use networked connections to shorten the time it takes to identify, and respond more efficiently and effectively to, public health perturbations.

Patty Mabry will complete the symposium by offering a glimpse of efforts underway to connect the biomedical research efforts supported by the National Institutes of Health (NIH) with the advanced computing initiatives sponsored by the National Science Foundation. Dr. Mabry, who works in the NIH Office of Behavioral and Social Science Research, is working with NIH leaders to improve knowledge management and to shorten time-to-discovery through cyberinfrastructure in the fields of biomedicine and public health.

References

- [1] B. J. Culliton, "Extracting Knowledge From Science: A Conversation With Elias Zerhouni," *Health Aff (Millwood)*, 2006.
- [2] B. W. Hesse, "Harnessing the power of an intelligent health environment in cancer control," *Stud Health Technol Inform*, vol. 118, pp. 159-76, 2005.
- [3] A. C. von Eschenbach, "A vision for the National Cancer Program in the United States," *Nat Rev Cancer*, vol. 4, pp. 820-8, 2004.
- [4] R. A. Hiatt and B. K. Rimer, "A new strategy for cancer control research," *Cancer Epidemiol Biomarkers Prev*, vol. 8, pp. 957-64, 1999.
- [5] American Cancer Society, "Cancer Facts and Figures, 2006," American Cancer Society, Atlanta, GA 2006.

U.S. National Science Foundation Program Managers: Perspectives on Sustaining Digital Government Research

Lawrence E. Brandt
Program Manager

Information & Intelligent Systems
Division

Computing & Information Sciences &
Engineering Directorate
National Science Foundation
4201 Wilson Blvd.
Arlington, Virginia 22230
703-292-8930
lbrandt@nsf.gov

Sylvia Spengler

Program Manager

Information & Intelligent Systems
Division

Computing & Information Sciences &
Engineering Directorate
National Science Foundation
4201 Wilson Blvd.
Arlington, Virginia 22230
703-292-8930
ssspengle@nsf.gov

1. Abstract

Two champions of digital government research will offer perspectives and ideas for helping to sustain digital government research at the U.S. National Science Foundation. There are changes in solicitations and new opportunities for researchers wishing to undertake research in the digital government domain.

The Information and Intelligent Systems Division at NSF is being reorganized into new clusters of research domains. There will most likely be a cluster tentatively entitled "Informatics and Information Integration (III). Members from the old program days incorporated into the new cluster include: *Digital Government (DG); Digital Libraries and Archives; Science and Engineering Informatics and Information Integration (SEI); and, Information, Data and Knowledge Management*.

Slicing the III cluster another way, there will be "*core*" research in these areas and "*contextual*" research. The latter is what SEI and DG have been doing, that is, bringing the CS research out of the lab and into various application areas (contexts). Proposers will be asked to identify their submissions as core or contextual.

Other clusters within the IIS Division are: *Human Centered Computing* - from Programs in Digital Society and Technologies, Human-Computer Interaction, and Universal Access; *Robust Intelligence* - from Programs in Artificial Intelligence and Cognitive Systems, Computer Vision, Human

Language and Communication, and Robotics

There will also be two Division-wide themes, *Human-Robot Interaction* and *Information Security and Privacy*.

The title of the Division-Wide call for proposals (encompassing all three clusters) is basically just the names of the three clusters combined. The IIS Division/NSF hope is to release the call in May, with proposals due in October.

2. General Terms

International, E-Government, Digital Government

3. Keywords

Science and Engineering Informatics and Information Integration, Digital Government, National Science Foundation

4. Speakers

Larry Brandt, Digital Government Program Manager

Sylvia Spengler, Science and Engineering Informatics and Information Integration Program Manager

Panel: Eco-Informatics and Decision Making

Managing Our Natural Resources

Judith Bayard Cushing
(Organizer)

Member of the Faculty

The Evergreen State College

Olympia, WA 98505

01-360-867-6652

judyc@evergreen.edu

Tyrone Wilson
(Moderator)

USGS Center for Biological Informatics

Reston, VA 20192

01-703 648-4164

tyrone_wilson@usgs.gov

Fred Martin

Senior Forest Biometrician

WA Dept. of Natural Resources

fred.martin@wadnr.gov

John Schnase

Goddard Space Flight Center

NASA

schnase@gsfc.nasa.gov

Sylvia Spengler

NSF Program Director

Information Integration and Informatics

sspengl@nsf.gov

Larry Sugerbaker

Vice President and CIO

NatureServe

Larry_Sugarbaker@natureserve.org

Theresa Pardo

Deputy Director, Center for Technology in Government

University at Albany, SUNY

tpardo@ctg.albany.edu

ABSTRACT

This panel responds to the December 2004 workshop on *Eco-Informatics and Decision Making* [1], which addressed how informatics tools can help with better management of natural resources and policy making. The workshop was jointly sponsored by the NSF, NBII, NASA, and EPA. Workshop participants recommended that informatics research in four IT areas be funded: modeling and simulation, data quality, information integration and ontologies, and social and human aspects. Additionally, they recommend that funding agencies provide infrastructure and some changes in funding habits to assure cycles of innovation in the domain were addressed. This panel brings issues raised in that workshop to the attention of digital government researchers.

Categories and Subject Descriptors

H.2.8 [Database Applications]: Scientific Applications.

Keywords

Information integration, knowledge management, modeling and simulation, data quality, web services, eco-informatics.

1. INTRODUCTION

The 1998 PCAST report characterized bio-informatics as a biology and CS/IT cross-discipline, recognized the biodiversity-ecosystem nexus as an information enterprise, and envisioned

analytical and synthetic capabilities among other foci in the next generation of NBII-2 information services [2]. Informatics tools for solving environmental challenges, e.g., global climate change, emerging diseases, decreasing biodiversity, and waning resources, are researched and developed under the *eco-informatics* rubric, but IT research supporting natural resource management remains largely absent from the academy.

In December 2004, the NSF, NBII, EPA, and NASA sponsored a workshop on Eco-Informatics and Decision Making [1], and At the workshop practitioners from government and non-government agencies, and university researchers from the computer, social, and ecological sciences, sought to address the gap described above. Participants found (IT) problems faced by resource managers in five categories: data presentation, data gaps, tools, indicators, and policy making and implementation.

As we go to press, no solicitation by funding agencies has as yet been published, but various NASA and NBII, and NSF cyber-infrastructure and digital government research efforts, now underway, address the above issues.

2. PROBLEM AND RESEARCH SPACES

Required information technology for eco-informatics policy-making is broad partially because political aspects necessitate equitable information access and mass customization, and distinguishing among measurements, indicators, and interpretations. Problems are categorized into five areas: 1) data presentation, 2) geographic data gaps, 3) tools, 4) indicators, and

5) policy and implementation problems. Four research areas were identified:

1) Modeling and simulation, including diverse model coupling, model design, values for diverse stakeholders, visualizations of model results, error and uncertainty representations, problems of large data sets, and open source modeling infrastructure.

2) Data quality problems, specifically how to determine and communicate uncertainty to decision-makers for studies integrating multiple data sources, and methods to mitigate error when creating and combining data sets and to associate error with alternative decisions.

3) Information integration, including mechanisms for reliable, transparent and authoritative data combination; integrating multiple ontologies; promoting document modeling; methods to evaluate the utility of both quantitative and qualitative data.

4) Social and human aspects of eco-informatics and policy including collaboration in IT tool development and information sharing, measuring success, determining appropriate institutional designs and incentives or disincentives; human-computer interaction; management practices, and data management training.

A special appendix of the workshop report defined “decision making” in the context of eco-informatics [3], and contains references that might be of value to researchers in this area.

3. THE DG.O PANEL

Dg.o panelists will present some current problems faced by natural resource managers, as well as current systems and current systems that aim alleviate those problems and research opportunities for those interested in this area.

Moderator Tyrone Wilson from the USGS National Biological Information Infrastructure will start the panel by describing some NBII efforts at helping provide tools for resource managers.

John Schnase will follow up the first presentation with an overview of NASA's efforts to provide data and tools that might be of use to natural resource managers, as well as some of the projects funded by NASA that might be relevant.

NatureServe is a consortium of 75 natural heritage programs spread over the U.S., Canada, and Latin America. Their traditional focus has been on at-risk species and ecosystems, and invasive species. They have a standard data model and information infrastructure to exchange and manage data from the member programs. Larry Sugarbaker will talk about how the nearly 90-person staff advances their data repository, and analysis and decision-making tools.

The Washington State Department of Natural Resources (DNR) manages several million acres of forested lands, both to support K-12 and in trust for future Washingtonians. Fred Martin is a senior biometrician for the DNR whose primary interest is in forest regeneration and succession, and providing field foresters with forest growth and development modeling tools. He will talk about problems that local forest managers face in making decisions about when and where to harvest timber annually on state-owned sites. He will also describe WA DNR and USFS tools currently available to those resource managers and talk about the shortcomings of those current tools.

Theresa Pardo will address why government resource managers might have difficulty using data sets and information systems to make management decisions.

Before opening the floor for general discussion, Sylvia Spengler will round out the discussion by suggesting ways in which researchers interested in this problem area might seek funding.

4. ACKNOWLEDGMENTS

We acknowledge the work of organizers of the Eco-Informatics for Decisions Makers Workshop: Frank Biasi (The Nature Conservancy), Larry Brandt (NSF), Judith Cushing (The Evergreen State College), Mike Frame (NBII/USGS), Valerie Gregg (then of the NSF), Eric Landis (private consultant), John Schnase (NASA), William Sonntag (EPA), Sylvia Spengler (NSF), Christina Vojta (USFS), and Tyrone Wilson (NBII/USGS), many of whom also served as report co-authors.

We thank the National Science Foundation (NSF IIS 0505790) for their funding of the 2004 Workshop and the USGS-NBII for publication of the Workshop Report.

5. REFERENCES

[1] The Eco-Informatics and Decision Making (BDEI3) workshop website includes all presentations, presenters and participants information, and the workshop final report.

<http://www.evergreen.edu/bdei>.

[2] PCAST. Panel on Biodiversity and Ecosystems. “Teaming with Life: Investing in Science to Understand and Use America’s Living Capital.” March 1998.

<http://www.nbii.gov/about/pubs/twl.pdf>.

[3] János Fülöp, David Roth, Charles Schweik. “Decision Making in the Context of Eco-informatics. Appendix 1 of the BDEI3 Final Report, pp. 21-26.

Sustaining An International DG/E-Gov Research Community

Sharon S. Dawes

Director

Center for Technology in Government
University at Albany/SUNY
187 Wolf Road, Suite 301
Albany, New York 12205
518- 442-3892
sdawes@ctg.albany.edu

Valerie J. Gregg

Assistant Director for Development
Digital Government Research Center
Information Sciences Institute
University of Southern California
3811 N. Fairfax Drive, Suite 200
Arlington, VA 22203
703-975-4777
vgregg@isi.edu

1. Abstract

This panel of experts will explore perspectives and opportunities for championing and sustaining an international digital government/e-government (DG/E-Gov) research community. The panel discussions will set the stage for the open international meeting that will develop future scenarios of e-government that will be held at dgo2006 on Wednesday afternoon (1:00 - 5:00 PM).

2. General Terms

International, E-Government, Digital Government

3. Keywords

Transnational, comparative, government, information society, e-gov, digital government

4. Background

Over the past decade, growing evidence demonstrates the emergence of a global field of inquiry at the intersection of government, society, and information and communication technologies. This domain is often characterized by "e-government," "e-governance," "information society," and other related terms. The term "digital government" encompasses this collection of research ideas in the United States. In Europe, the European Commission, as part of its Information Society Technologies (IST) program, sponsors an ambitious e-government research program. At the same time, the research councils of individual European states support comparable research programs within their borders. Similar efforts are established or emerging in Canada, Australia, India, the Pacific Rim, Latin America, and Africa. International organizations such as the United Nations and World Bank support e-government development and are also becoming interested in associated research.

Because of the relative newness of the DG/E-Gov field, there is insufficient interaction among researchers in different countries

compared to what one finds in more established scientific disciplines. As this is a relatively new domain of inquiry, it involves multiple disciplines (a challenge within a single country, let alone internationally) and there are very few support mechanisms and forums to engage DG/E-Gov researchers with their peers working in this domain around the globe. Furthermore, once a potential collaboration starts that could lead to joint research efforts, it is logically and financially difficult to sustain it to the point of joint research proposals and reliably funded projects. Consequently, comparative and transnational issues in DG/E-Gov, which are of growing importance in an increasingly networked world, are not receiving the attention they deserve.

This panel highlights efforts now underway in different parts of the world to encourage and sustain digital government researchers and to initiate research projects that address international problems and comparative questions.

5. Speakers

- Dr. Sharon S. Dawes, Center for Technology in Government, SUNY Albany
- Dr. Roland Traunmuller, University of Linz, Austria
- Dr. Tomasz Janowski, United Nations University
- Dr. Atreyi Kankanhalli, National University of Singapore
- Dr. Maria Wimmer, University of Koblenz-Landau, Germany

Symposium Summary

Dr. Dawes will describe a four-year NSF award for sustaining an international digital/e-government research community. The streams of work include a reconnaissance study of the status of worldwide digital government research projects and institutions supporting research in this domain of inquiry; a plan for forming and supporting international working groups; and, an annual research institute for PhD. students.

Dr. Traunmuller will describe growth and history of the European E-Gov research conference held in conjunction with the annual DEXA conference. He will offer perspectives on growing and sustaining an international digital/e-government research community.

Dr. Tomasz Janowski will describe the United Nations E-government program. He will offer perspectives and describe opportunities on how the global digital/e-government research community can become involved.

Dr. Atreyi Kankanhalli will describe the digital/e-government programs and projects in Singapore. She will provide guidance on Pacific-Rim opportunities for digital/e-government research community.

Dr. Maria Wimmer will discuss the European Union's Road Map 2020 project for defining future e-government scenarios and other relevant EU projects related to global digital/e-government community building.

RESEARCH PRESENTATIONS

SESSION 1A

DATA MINING

Moderator

Jamie Callan, Carnegie Mellon University, USA

Titles and Authors

Spatiotemporal Analysis of 9-1-1 Call Stream Data
Jasso, Hector; Hodgkiss, William; Fountain, Tony; Baru, Chaitan; Reich, Don; Warner, Kurt

A Temporal Based Forensic Analysis of Electronic Communication
Stolfo, Salvatore J.; Hershkop, Shlomo; Creamer, German

Using Semantic Components to Facilitate Access to Domain-Specific Documents in Government Settings
Price, Susan; Delcambre, Lois; Nielsen, Marianne Lykke; Tolle, Timothy; Luk, Vibeke;
Weaver, Mathew

Spatiotemporal Analysis of 9-1-1 Call Stream Data

Hector Jasso
Tony Fountain
Chaitan Baru
San Diego Supercomputer Center
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0505
hjasso@sdsc.edu, fountain@sdsc.edu,
baru@sdsc.edu

William Hodgkiss
Scripps Institution of Oceanography
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0701
whodgkiss@ucsd.edu

Don Reich, Kurt Warner
Public Safety Network
Santa Barbara, CA
donreich@donreich.com,
kurt@kurtewarner.com

ABSTRACT

Currently, archival 9-1-1 call stream data is used mainly for administrative purposes. We present spatiotemporal analysis of thirteen months worth of call stream data for the purpose of illustrating how this data might be used for enhancing emergency response in the State of California. An analysis of the data shows regularity in the 9-1-1 call volume which can facilitate the automatic detection of abnormally high call volumes that are associated with environmental, medical emergency, and law enforcement events. Thus, this is a first step towards the detection of unusual trends that could indicate widely spread events that require response beyond that of isolated incidents.

Keywords

Emergency response, public safety, telephone 911.

1. INTRODUCTION

Although 9-1-1 call activity data is collected at the State level, this information is currently used only to generate monthly reports, mainly for allocation of funding and staffing of 9-1-1 call centers. First responders could utilize this information to plan their local response and at the State level this information also could be used to generate immediate and dynamic information on the human impact of moderate to large-scale disaster events and facilitate timely resource allocation. Thus, this data can form the basis of an *early warning system* to assist in the regional analysis of developing emergency situations along with the allocation of resources for emergency response.

For this purpose, we explored the viability of using 9-1-1 call stream data as a proxy for large, atypical events, and as a first sensor for information related to the magnitude and location of such events. Preliminary analysis of thirteen months of collected data showed a normally distributed daily call volume and a well-defined pattern in the hourly call volume. Additionally, spatiotemporal analysis of 9-1-1 call activity during a particular emergency event showed a localized burst of above-average 9-1-1 call activity shortly after the event and around the event's location. This first analysis thus supports the idea of using 9-1-1 call volume information for detecting emergency events.

2. COLLECTION OF 9-1-1 CALL STREAM DATA

2.1 Data Collection

For the purpose of this research, archival 9-1-1 call stream data is being used. In generating a database from this archive suitable for analysis, a variety of corrections have been made (e.g. time stamp corrections) as well as the conversion from street addresses to latitude/longitude. The information available includes: *Call date*, *Call time*, *Call location* (latitude/longitude), *Public Safety Answering Point (PSAP) identification number*, *Length of time for call to be answered*, *If call has been abandoned*, *Duration of the call*, and *Phone type* (e.g. business, residence, wireless).

3. 9-1-1 CALL STREAM DATA ANALYSIS

Thirteen months of data (September 1, 2004 to September 30, 2005) for the San Francisco Bay Area have been collected with a total of 1,856,170 calls corresponding to 66 PSAPs.

Figure 1 compares hourly call volume for weekday and weekend days for the San Francisco Combined Emergency Communications Center (CECC) PSAP. During the early morning hours the number of calls is lowest and increases during the day, peaking at around 4PM. No major differences were found between individual weekdays, except for Friday evenings, which had a similar call volume to Saturday evenings.

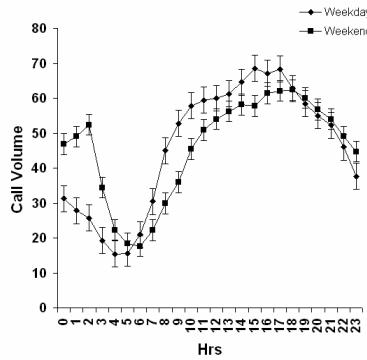


Figure 1. Average hourly call volume for the San Francisco Combined Emergency Communications Center (CECC). Bars indicate standard errors. Both weekdays and weekend days are shown.

4. STEPS TOWARDS IMPLEMENTING AN ALARM SYSTEM

4.1 Burst Detection

Given the regularity of call volume, bursts in call volume can be taken as possible emergency events. For this, an individual call can be represented as a single data point in 3-dimensional space (longitude, latitude, time). Simple statistical thresholds can be computed in a manner analogous to conventional process control. In this approach, a measure of the event occurrence is calculated based on statistics over a fixed space and a fixed time window. For example, an alert level can be calculated as 2 or 3 standard deviations from the mean over a given spatial neighborhood over a given window of time.

This approach, however, should be adapted to the various spatial and temporal scales and call activity variations over various temporal scales, from daily to seasonally. Furthermore, the normal call level expectations are dependent on significant, but normal, events, for example, the 4th of July holiday. A similar complication arises from the multiple spatial scales. For example, significant local variation at the neighborhood level would be lost in an aggregate measure at the state level. A multi-resolution analysis approach in both the spatial and temporal dimensions should be employed to address this issue.

Finally, although the process control approach can capture certain classes of novel events, namely those with frequency counts that exceed expectations, they are inadequate to capture more complex relations between call events. For example, call stream data for significant events might exhibit a clustering behavior, often radiating along lines corresponding to civil infrastructure (highways, neighborhoods, etc).

4.2 Example: Walnut Creek Explosion

In this section we describe the spatiotemporal characteristics of a particular emergency event which generated a clearly detectable burst in call volume. Figure 2.a shows call activity after an explosion in Walnut Creek, CA, on November 9, 2004. Figure 2.b shows how, shortly after the event, a number of 9-1-1 calls were made, clustered geographically around the event. The number of calls after the event by far exceeded the normal call activity for other times during the same day.

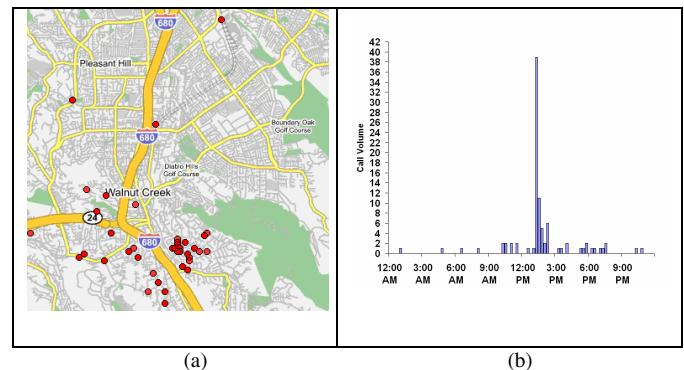


Figure 2 a) Warehouse explosion in Walnut Creek, CA. Warehouse near the tight cluster of calls. Spatial distribution of call activity during the event (1PM to 2PM) is shown. Event happened at 1:30PM. b) Calls for the Walnut Creek Police Department PSAP during Nov 9, 2004, shown in 15-minute bins. The Walnut Creek explosion at 1:30 PM generated a call volume well above the average, and extending about 1 hour after the event.

5. CONCLUSION

We presented the first steps in the use of 9-1-1 call stream data for detection of spatiotemporal patterns corresponding to emergency events. Although burst detection can be a difficult task given the complexity of the underlying phenomena behind 9-1-1 calls, in the event presented the call volume was considerably above average within time and space. This points to the possibility of emergency event detection based on outlier detection.

Trend analysis of 9-1-1 call stream data could be used for improvement of the overall emergency response system by facilitating dynamic resource allocation in order to provide timely assessment of the location and magnitude of incipient disasters. Eventually, the results of such analyses can be used to feed a “reverse 9-1-1” system, i.e. to broadcast warnings to the public about emergency situations in a region. Through more rapid identification and localization of a significant event, emergency response services can be brought to bear earlier and more efficiently thus leading to a more rapid and smooth recovery along with a corresponding reduction in cost.

ACKNOWLEDGMENTS

This research was supported by the National Science Foundation under grant IIS-0429448. Jim Watkins and Megan Glasscock from the California Governor’s Office of Emergency Services participated in the original development of this project.

A Temporal Based Forensic Analysis of Electronic Communication

Salvatore J. Stolfo

Columbia University
500 West 120th St
New York, NY 10027
1-212-939-7080

sal@cs.columbia.edu

Germán Creamer

Columbia University
500 West 120th St
New York, NY 10027
1-646-775-6068

gcreamer@cs.columbia.edu

Shlomo Hershkop

Columbia University
500 West 120th St
New York, NY 10027
1-646-775-6041

shlomo@cs.columbia.edu

ABSTRACT

Previous work [1] reported on our research in developing a data mining environment for analyzing email communication data. In this paper, we describe our extensions to EMT for applying forensic discovery over temporal email data. The goal is to produce a semi-automatic system to aid in evidence discovery and a host of other applications. We describe our research on profile stability, temporal search and clustering, and new social network dynamic algorithms.

Categories and Subject Descriptors

H.4 Information Systems, H.4.3 Communication Applications, Electronic mail and Information Browsers

General Terms

Algorithms, Theory.

Keywords

Social Networks, Email Mining, Histograms, Search, Email Visualizations.

1. INTRODUCTION

The analysis of *email flows* to and from a user's email account(s) reveals a tremendous amount of information about a person's interests, activities and behaviors that cannot be derived alone from content analyses of individual emails. In order to develop email forensic tools, we focus on several aspects of email flows: evidence discovery; profiling, interactions, and content communication.

The behavior analysis of individual users over time is one way to start an investigation dealing with email. By computing individual behavior over time, we can more accurately study changes in behavior for both individual users and groups of users, and find interesting points in time and discover influences of interest.

The second aspect of our work is to study how to apply forensics to social group interactions over time. Our approach is to visually allow the group features to be adjusted based on features which are important to the end user analyst.

Last, locating interesting emails through searching is a non-trivial task since the search terms are not always clear ahead of time. In a digital evidence framework, one would prefer a system of being able to search through emails using time as a basis but also to include content similarity across all messages. We have augmented EMT's current search capabilities to automatically expand and visualize search parameters based on time and word relationships.

2. EMT

The Email Mining Toolkit (EMT) has been in development at Columbia University since 2001 and featured in numerous publications. It is a Java-based data mining environment for large email collections focused on automatically extracting patterns of users, social group interactions, and attachment level analysis of email communication. EMT has been downloaded by dozens of organizations.

3. Profiling Account Use

User level analysis is based on profiling individual email accounts over time. Specific features are used to create a profile over a period of time ('baseline normal') and compare future behavior to this profile. This static approach has been useful in locating similar behaving accounts in a large email collection. The reality is that true behavior is dynamic over time, and a profile representation should be augmented by adjusting it over time to detect usage changes within the same account.

3.1 Rolling Histogram

A rolling histogram is a dynamic per user profile computed over time. We can measure profile stability over time, by sliding a set window over the behavior data, and computing a similarity measure for two adjacent time periods. For example, given a year of email data we can compute weekly or monthly profiles and compare the behavior of the account within these small increments.

We define account stability to be distance between periods under some average threshold. One can view the changes in histogram distance scores over time using the interface. Alerts are generated for those periods which signify unusual behavior allowing time periods to be further investigated (i.e. all emails can be examined over a specific period to see what triggered the alert).

4. Social Mining

EMT allows social networks to be extracted both on a per user basis and per enclave basis. We have extended the basic feature with the ability to generate clusters among related email accounts between senders and recipients (figure 1). It calculates the shortest path length (average distance) from a specific vertex to all vertices in the graph, and the cluster coefficient for each vertex. For every user, these indicators should be stable among a community of users. Therefore, outlier values may indicate suspicious patterns. This module is also able to discover communities of users based on voltage drops or weaker connections across networks (figure 2).

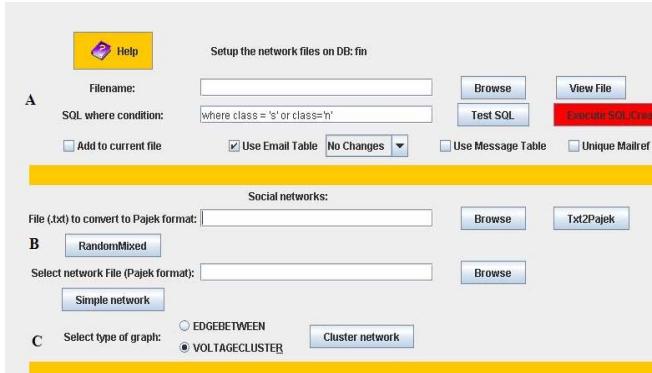


Figure 1 - Email social network window. (A) Network file can be generated with sender and recipient email accounts. (B) Text files can be converted to Pajek format. (C) Several types of social networks and clusters can be generated with email or external files.

The social network module has the capacity to process different social structures from external sources or email accounts. It has the flexibility to receive and generate text files with the structure of a network and convert it onto the Pajek format. The Pajek format is a well-known format in the social network literature that is used to generate new networks or clusters.

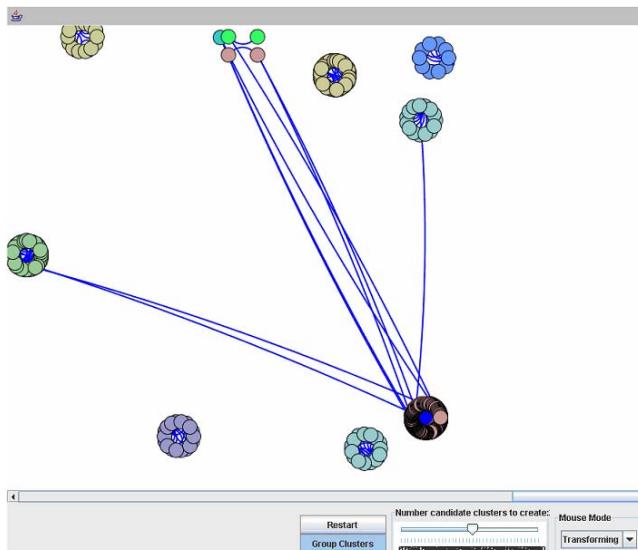


Figure 2 Clusters of users created according to the quality of connection between networks.

5. Search Modules

Unusual communication patterns are a focus of this research, but equally as important is anomalous communication content which would be of interest to a forensic investigator.

We have been exploring extending the basic search functionality with a two step discovery process. We pre-compute a search index over all emails. When a search term is executed, we highlight a time line calendar with all relevant results and display a visualization of the contextual relationship of the

number of emails per calendar entry of the results. We then use a thesaurus to grab similar words and unionize the results of those searches to find similar email communication.

6. Progress Report

In our last report [1] we noted our efforts in establishing a trial run of EMT with the New York Police Department. Unfortunately, our contact retired before being able to start the pilot study. Currently EMT is being evaluated by a number of commercial and research entities and has been offered to NARA in their new digital archive initiative managed by Lockheed Martin.

7. Future Directions

The flexibility of the social network module is useful for other security related tasks that may or may not be able to be detected through emails.

Our most recent emphasis is on extracting a user's social network given the user's email inbox [2] or on visualizing the social networks generated by email interactions [3], we use average distances and cluster coefficients to identify unusual relationships among members of different groups.

Some of the areas that this module can be used for include the creation of a "Chinese wall" among financial analysts and investment bankers. The social network module might be able to detect unauthorized activity between these groups. In addition it could also be applied to detecting links or conflict of interests among directors, and financial analysts that may bias financial reports. These indicators could also be used as part of a financial information system to predict stock prices or corporate performance.

We would like to thank both the National Science Foundation and DARPA for funding this work. In particular NSF grant - Email Mining Toolkit Supporting Law Enforcement Forensic Analysis from the Digital Government research program, No. 0429323.

8. REFERENCES

- [1] Hershkop, S. *Behavior-based Email Analysis with Application to Spam Detection*. Ph.D. Thesis, Columbia University, New York, NY, 2005.
- [2] Culotta, A., Bekkerman, R. and McCallum, A. *Extracting Social Networks and Contact Information from Email and the Web*. First Conference on Email and Anti-Spam (CEAS), Mountain View CA, 2004.
- [3] Boyd D. and Potter, J. *Social Network Fragments: an interactive tool for exploring digital social connections*. SIGGRAPH, San Diego CA, 2003.
- [4] Stolfo, S. *Email Mining Toolkit Supporting Law Enforcement Forensic Analyses NSF Final Report*. DG.o 2005 Atlanta, GA. May 2005.

Using Semantic Components to Facilitate Access to Domain-Specific Documents in Government Settings

¹Susan Price, ¹Lois Delcambre, ²Marianne Lykke Nielsen, ³Timothy Tolle,
⁴Vibeke Luk, ⁵Mathew Weaver

¹Computer Science Dept.
Portland State University
{prices, lmd}@cs.pdx.edu

²Royal School of Library & Information Sciences, Aalborg, Denmark
mln@db.dk

³Consultant
Vancouver, WA
TimTolle@aol.com

⁴Sundhed.dk
Copenhagen, Denmark
vlu@sundhed.dk

⁵Consultant
Paradise, Utah
mweaver@cs.pdx.edu

ABSTRACT

We introduce the notion of *semantic components* that occur in domain-specific documents, discuss use of semantic components to improve document retrieval in a domain-specific collection, and present the results of two case studies.

Categories and Subject Descriptors

H.3.1 [Information Storage and Retrieval]:Content Analysis and Indexing H.3.3 - *Indexing methods* [Information Storage and Retrieval]: Information Search and Retrieval - *Search process*
H.3.7 [Information Storage and Retrieval]: Digital Libraries.

1. INTRODUCTION

Our aim is to support the information needs of domain-specific users, such as natural resource managers making decisions about public lands and healthcare practitioners making decisions about caring for individual patients. Such users are professionals whose specialized information needs are connected to specific work tasks, specific document types, and specific perspectives on the topics of interest. Since they are often making decisions in limited time or engaging in focused (rather than open-ended) research tasks, we focus on improving search precision.

We hypothesize that users' knowledge about document types, and about how knowledge is organized and expressed by other domain experts within documents, can be leveraged to facilitate more precise searches. Our idea is to define *semantic components* as a new model for describing document content. This allows users to express information needs with respect to semantic components. For example, a physician seeing a patient with puzzling symptoms might be interested only in documents with information about *diagnosis* of migraine headaches, not about *treatment* or *prevention*. A natural resource manager may wish to find documents where "spotted owl" appears in a discussion of the *alternatives* in an Environmental Analysis, but

not care about occurrences of the term in other contexts. We build on theories and other research about document structures and genres [1-4].

We investigated documents in two domains where we identified document classes and corresponding semantic components. We discuss how we can apply this knowledge about document classes and semantic components for improved information retrieval.

2. DOMAINS AND PARTNERS

Our research takes place in the context of two related digital government projects. In our current international project, we have a new partner, sundhed.dk, the operational, national, web-based health portal in Denmark. The health portal contains over 20,000 documents that provide information about health and the healthcare system to healthcare professionals and to the public.¹ We are also continuing a partnership with Region 6 of the USDA² Forest Service (FS). The *Forest Project*³ has resulted in a prototype digital library system [5, 6] that is intended to contain documents produced and used by government agencies that manage and make decisions about public lands.

3. SEMANTIC COMPONENTS

We define a *semantic component instance* as section(s) of text, variable in length, that contain information about a particular aspect of an instance of a concept that is important in a domain. A *semantic component* is the type (or label for a type) for the instances corresponding to a particular aspect. We postulate that it is possible to identify classes of documents that contain similar collections of semantic components. For example, documents about diseases often discuss treatment options and possible complications of the disease. Semantic components in such documents include *treatment* and *complications*; the semantic component instances are the sections of text that describe treatment and complications. Similarly, environmental analyses will include the semantic components *alternatives* and

¹ Vibeke Luk is the primary point of contact with the Danish Health Portal.

² United States Department of Agriculture

³ "Harvesting Information to Sustain Our Forest" NSF Project Number EIA9983518

environmental consequences. While documents are often organized around semantic components, the semantic components may or may not be explicitly indicated with structural elements, and they need not be confined to a single location in the document.

We believe semantic components may be useful for retrieval in two general ways: relevance assessment (helping the user decide which documents to view), and search specification. For example, a patient scheduled for knee surgery might wonder what he can eat or drink the morning of the procedure. A search on “knee surgery” might yield too many documents, but knowing that a document contains the semantic component *preparation* would identify it as potentially useful. A different searcher might be interested in the term “radiation,” but only if it appears in the *etiology* component of a document about cancer.

4. DOCUMENT ANALYSIS

We analyzed a random sample of documents from each domain to verify the feasibility of identifying document classes and semantic components. The health portal sample was taken from all documents available through the advanced searching interface. The forest portal sample was limited to two document types, Environmental Analysis (EA) and Decision Notice (DN). We outlined each document to summarize its content and made a preliminary list of the types of information present.

We classified the health documents according to *intended audience* (health professionals or patients); whether documents were *clinical*; and whether documents were *region-specific*. We further subclassified documents for health professionals by the type of primary topic: a clinical problem (such as a disease), or a test or procedure (such as a laboratory test). We also subdivided documents for patients according to primary topic: clinical problem, procedure or test, or wellness and health maintenance.

We further analyzed two classes of health documents: *clinical documents written for health professionals about problems* (ClinProfProb) and *clinical documents written for patients describing procedures and tests* (ClinPatProc). We identified ten semantic components in ClinProfProb documents: *evaluation, therapy, management guidelines, referral guidelines, prevention, risk factors, prognosis, etiology, associated conditions, and epidemiology*. In the ClinPatProc documents we identified six semantic components: *preparation, practical details, description of procedure, risks and complications, aftercare, and where to direct questions*.

Within each class, neither presence nor location of a particular semantic component in a document was predictable. However, each document in a class contained one or more of the semantic components listed, and the semantic component instances comprised much of the document content. Structural elements to aid identification of components were present in some cases but often were absent.

We classified the Forest Portal documents according to document type (EA or DN) and identified semantic components for each. EAs and DNs are both mandated by the National Environmental Protection Act (NEPA) to document significant projects by the FS and other agencies. Although length and

detail varies, characteristic content and structure makes EAs and DNs easy to identify. NEPA guidelines and templates specify content elements that are essentially semantic components. For example, DNs contain: *Background, Decision, Alternatives, Rationale, Mitigation measures, Public involvement, FONSI, Findings required by other laws, Implementation date, Administrative review, Contact person, and Responsible official*. The DNs and EAs contained most, if not all, the semantic components for the document class. Semantic component instances often (but not always) corresponded with section headings, but the text of the headings varied. Because many documents are quite long, and most documents of the same type will contain the same components, being able to search the content within semantic components could be very useful.

5. CONCLUSIONS AND FUTURE WORK

In two very different document collections we used domain knowledge to classify documents and to identify semantic components in classes of documents that we believe will be useful for retrieval. Future work includes testing the usefulness of semantic components to searchers and testing the accuracy and consistency of identifying semantic components. Identifying semantic components can be manual, automatic, or semiautomatic, and could be combined with either keyword assignment or full text indexing.

6. ACKNOWLEDGMENTS

This work is supported in part by the National Science Foundation, grant number 0514238. Any opinions, findings, conclusions, or recommendations expressed here are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

7. REFERENCES

- [1] Moens, M. Automatic indexing and abstracting of document texts. Kluwer Academic Publishers: Dorrecht, The Netherlands, 2000.
- [2] Hearst, M. and Plaunt, C. Subtopic structuring for full length document access. In Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 59-68, Pittsburgh, PA, 1993.
- [3] Freund, L., Toms, E. G. and Clarke, C. L. Modeling task-genre relationships for IR in the workplace. In Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval SIGIR '05, Salvador, Brazil, August 2005.
- [4] Freund, L., Toms, E. G. and Waterhouse, J. Modeling the information behaviour of software engineers using a work-task framework. In ASIS&T '05 Proceedings of the 68th Annual Meeting, Charlotte, NC, October 28-Nov2, 2005.
- [5] Weaver, M., Delcambre, L. and Tolle, T. Metadata++: A Scalable Hierarchical Framework for Digital Libraries. In Proceedings of the ICADL 2003, p. 696, 2003.
- [6] Weaver, M. Enhancing a domain-specific digital library with metadata based on hierarchical controlled vocabularies. PhD Dissertation, Oregon Health & Science University, 2005.

SESSION 1B

VOTING

Moderator

James P. Zappen, Rensselaer Polytechnic Institute, USA

Titles and Authors

Political E-Identity: Campaign Funding Data and Beyond
Wolber, David

A Project to Assess Voting Technology and Ballot Design
Traugott, Michael W.; Herrnson, Paul S.; Niemi, Richard G.

E-voting in the 2005 Local Elections in Estonia and the broader impact for future e-voting projects
Trechsel, Alexander H.; Breuer, Fabian

Political E-Identity: Campaign Funding Data and Beyond

David Wolber
University of San Francisco
2130 Fulton Street
San Francisco, CA., 94117
(415) 422-6451
wolber@usfca.edu

ABSTRACT

This paper describes a joint university-government effort to develop software that helps journalists, voters, and watchdog organizations visualize campaign funding data in San Francisco. The paper also presents broader plans for constructing comprehensive electronic identities for politicians, and describes how emerging trends in on-line information systems can be leveraged. The newest version of the software is at <http://www.whosfundingwhom.org>.

Categories and Subject Descriptors

K.4.1 [Computers and Society]: Public Policy Issues

General Terms

Design, Human Factors, Standardization

Keywords

Campaigns, Visualization, Public Participation, Semantic Web

1. INTRODUCTION

There has been little progress in campaign finance reform and public disclosure since Roosevelt spearheaded the Publicity Act of 1910 [8]. Computers have helped-- San Francisco and New York initiated the first required on-line filing systems in 1993[4] and there are now megabytes of data existing in thousands of databases across the country. Unfortunately, there is a lack of sufficient software for viewing that data. Existing software leaves journalists and voters to perform the virtual equivalent of rummaging through file cabinets to discover the web of financial relationships that control our elections and country.

The inaccessibility leads to: 1) Voters being ill-informed concerning who is funding candidates, 2) Journalists and investigators spending weeks uncovering information that could be at their finger tips, 3) Campaign ethics commission administrators investigating only the most egregious of filing violations. The end-result is less transparent campaigns and more

corporate influence handcuffing our leaders.

The lack of sufficient visualization software can be attributed to the shortage of funds for such projects in local and state governments, as well as the high cost of developing software. Most city information technology departments are overwhelmed with the challenge of making government services available on-line, and have thus far focused on implementing on-line forms and rudimentary viewing systems.

The Transparency in Government project at the University of San Francisco (USF) is an example of how universities can help fill this void. Working directly with the San Francisco City Ethics Commission directors and staff, USF students have created a software tool for viewing campaign finance data that is filed in San Francisco. The software periodically downloads the raw data from the Ethics Commission and builds graphs and web forms that make it easy for users to follow campaign funding trails.

The current version of the software, found at <http://www.whosfundingwhom.org> (see Figure 1), is the first step in a larger goal of providing comprehensive electronic identities for our leaders and those running for office. In this paper, we describe the current status of the software and the issues involved, then discuss strategies for extending it based on the semantic web and social software.

2. PROJECT ORIGINS

Through its Ethics commission (www.sfgov.org/ethics), San Francisco is one of the leaders in on-line filing—they enacted the first mandatory electronic filing bill and built the world's first on-line campaign finance database [1].

In terms of campaign finance data, San Francisco, like most municipalities, has henceforth focused on the input-side—replacing the old paper forms with on-line equivalents as required by newly enacted laws. Less attention has been given to output—allowing the public to easily visualize the data. Currently, a user can search the resulting database to retrieve forms and summaries. Though it is more than what most local municipalities provide, the site is not ideal for analyzing data—it does not allow a user to quickly view the key officers and candidates, navigate the money trails inherent to campaigns, or provide a graph view of multiple entities and funding relationships. As a simple example, when the user views the entities who have funded Mayor Newsom, the user cannot then click on them to see the entities that have funded the entities that fund Newsom.

Aware of the potential for this on-line data, as well as the lack of resources available to the Ethics Commission, Commissioner Joe

The screenshot shows a Mozilla Firefox browser window displaying the "OfficerList - Mozilla Firefox" page. The URL is http://zen.cs.usfca.edu/TGPOfficerView_2_1/OfficerList.aspx. The page title is "Who's Funding Who in San Francisco 2005" with the subtitle "making information visible to the common citizen...". The page features a green header with the USF logo and navigation links for "Current Officers", "Search", and "About". Below the header is a table with columns: "Office", "Current Office Holder", "Committee 1", "Committee 2", and "More". The table lists campaign contributions for various city offices:

Office	Current Office Holder	Committee 1	Committee 2	More
Mayor	Gavin Newsom Contributions Received: \$0.00 Contributions Made: \$0.00 Expenses: \$0.00	Newsom for Mayor Contributions Received: \$347,634.68 Contributions Made: \$0.00 Expenses: \$478,009.00		
District Attorney	Kamala Harris Contributions Received: \$0.00 Contributions Made: \$350.00 Expenses: \$0.00	Kamala Harris for Distric... Contributions Received: \$0.00 Contributions Made: \$0.00 Expenses: \$0.00		
Board of Supervisors: District 1	Jake McGoldrick Contributions Received: \$0.00 Contributions Made: \$0.00 Expenses: \$0.00	Committee to ReElect Sup... Contributions Received: \$0.00 Contributions Made: \$0.00 Expenses: \$0.00	Jake McGoldrick for Super... Contributions Received: \$0.00 Contributions Made: \$0.00 Expenses: \$0.00	Re-Elect Jake McG RE-ELECT SUPER
Board of Supervisors: District 2	Michela Alico-Pier Contributions Received: \$0.00 Contributions Made: \$500.00 Expenses: \$0.00	Michela Alico-Pier for S... Contributions Received: \$0.00 Contributions Made: \$0.00 Expenses: \$0.00		
Board of Supervisors: District 3	Aaron Peskin Contributions Received: \$0.00 Contributions Made: \$500.00 Expenses: \$0.00	Aaron Peskin for Supervis... Contributions Received: \$5,250.00 Contributions Made: \$0.00 Expenses: \$6,355.00		
Board of Supervisors: District 4	Fiona Ma Contributions Received: \$0.00 Contributions Made: \$1,050.00 Expenses: \$0.00	FIONA MA EXPLORATORY COMM... Contributions Received: \$0.00 Contributions Made: \$0.00 Expenses: \$0.00	FIONA MA FOR ASSEMBLY 200... Contributions Received: \$0.00 Contributions Made: \$0.00 Expenses: \$0.00	Fiona Ma for Super
Board of Supervisors: District 5	Ross Mirkarimi Contributions Received: \$0.00 Contributions Made: \$0.00 Expenses: \$0.00	Ross Mirkarimi for Superv... Contributions Received: \$1,500.00 Contributions Made: \$0.00 Expenses: \$1,601.71		
Board of Supervisors: District 6	Chris Daly Contributions Received: \$0.00 Contributions Made: \$100.00 Expenses: \$0.00	Campaign to Elect Chris D... Contributions Received: \$0.00 Contributions Made: \$0.00 Expenses: \$0.00		

Figure 1. The main page of whosfunding whom.org.

Lynn approached representatives from the Leo. T. McCarthy Center for Public Service and the Common Good at USF (<http://mccarthycenter.usfca.edu>). Under the direction of Patrick Murphy, the center is a focal point of the university's service learning initiative and the university's mission of educating hearts and minds. Commissioner Lynn's idea was met with enthusiasm from Murphy, the Center's namesake, Leo McCarthy, as well as former mayor of San Francisco Art Agnos, who serves on the Center's board.

The missing link was the connection to technology. Centers like the McCarthy Center, and public service efforts in general, have traditionally been concerned with humanities—sending students to work on political campaigns or inner-city soup-kitchens.

Fortunately, USF has emphasized service learning in the sciences as well, perhaps most prominently in the computer science department. The McCarthy center had already helped fund the department's Community Connections effort (www.usfca.edu/cc), which regularly sends students into inner city computer centers to provide information technology services and annually sends a group to Peru to build and maintain computer labs at needy schools.

Whereas the previous Community Connections projects focused on system administration tasks, this one would focus on software development, which is the primary focus of the computer science curriculum and more in-line with what most computer science students will end up doing for their careers. The department embraced the idea put forth by Lynn, Murphy, McCarthy and Agnos, and a collaborative effort was born.

Beginning in August of 2004, five USF students and one professor began development of the software. The key challenge was determining the specifications for the project, including gaining an understanding of the funding data and all of its complexities. This domain analysis was only possible through the efforts of the Ethics Commission directors and staff, most notably directors Mabel Ng and John St. Croix, as well as Oliver Luby, the staff member with the most expertise in actually using the funding data to uncover filing inconsistencies.

Through an iterative development cycle, analyzing top federal data sites like www.opensecrets.org and the Federal Elections Commission site (<http://www.fec.gov/disclosure.shtml>), and continual consulting with those at the Ethics commission, the

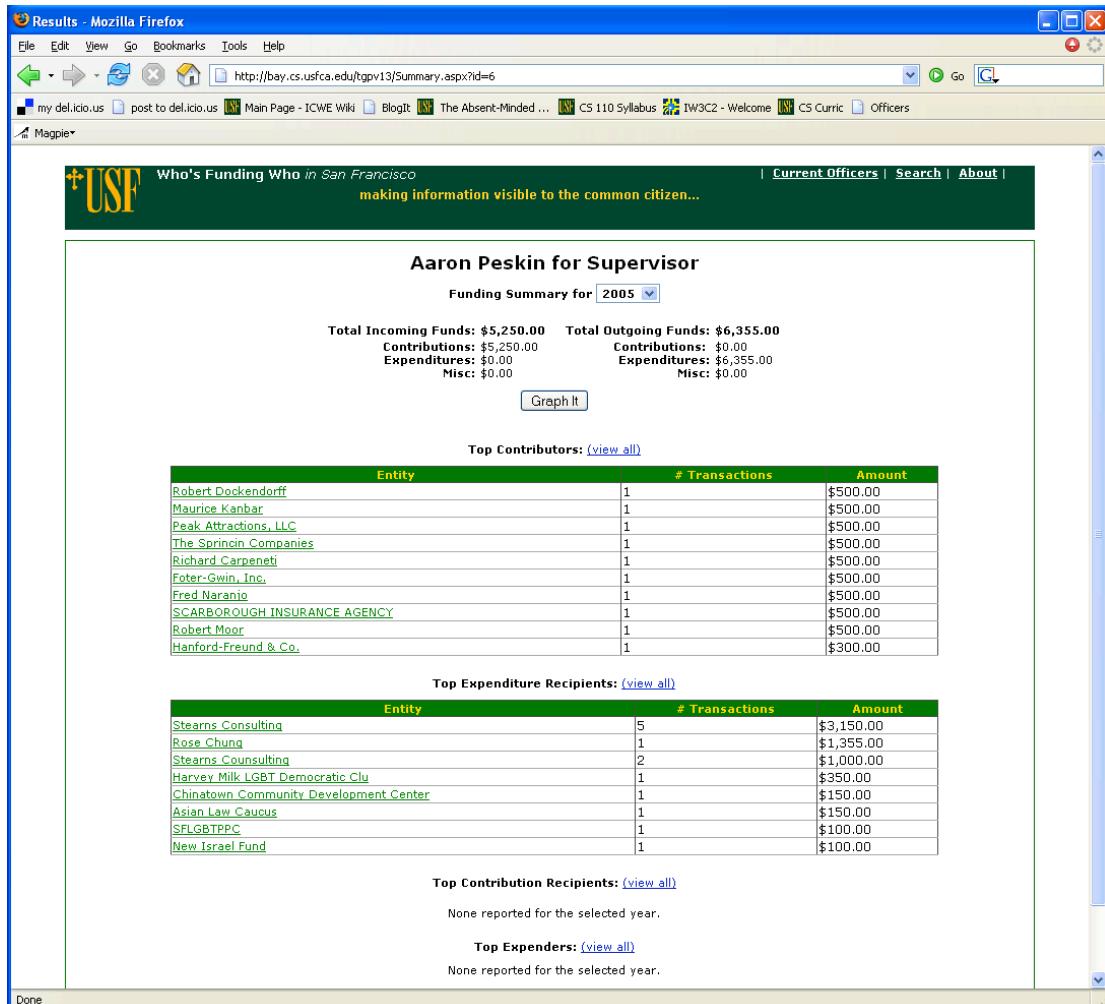


Figure 2. Individual Summary Page for Supervisor Aaron Peskin.

team completed a working version of the funding data visualization software, as well as the city's first on-line lobbyist filing system. The campaign funding site, which is the focus of this paper, is publicly available at <http://whosfundingwhom.org>, while the lobbyist software is now undergoing pilot testing with a selected group of lobbyists. Nine different students have participated in the project over the year and half of its existence, funded in part by grants from the McCarthy Center and the City of San Francisco.

3. SYSTEM OVERVIEW

The campaign funding visualization software takes raw data supplied from the Ethics commission, builds a relational database from it, then displays the data in forms that allows a user to easily view campaign activity and funding trails. The system shows only campaign funding data filed by political committees in San Francisco.

The key goal is to help users do what Deepthroat suggested to Woodward and Bernstein in the Watergate investigation: "follow the money". The system provides two methods of navigating such

trails: a table-based method that displays the data for a particular entity and allows quick navigation to the page for another entity (see Figure 2), and a graph view that provides a birds-eye view of multiple entities and chained relationships (see Figure 3).

3.1 Where do Users Start?

In the original Ethics Commission viewer, the landing page was a search page in which the user could enter a name and get a listing of the forms filed relating to that person or committee. Such a search page is necessary, but it assumes that the user is looking for a particular entity and has knowledge about the key players in the city's government. For most users, it is not an ideal introduction to the system.

A key innovation in the whosfundingwhom.org software is the introduction of a landing page which is a list of the key offices in San Francisco, including the Mayor, the District Attorney, and the Board of Supervisors (see Figure 1). This page gives the user a birds-eye-view of the politicians in San Francisco and quick summaries of the funding records for the politicians and their key committees.

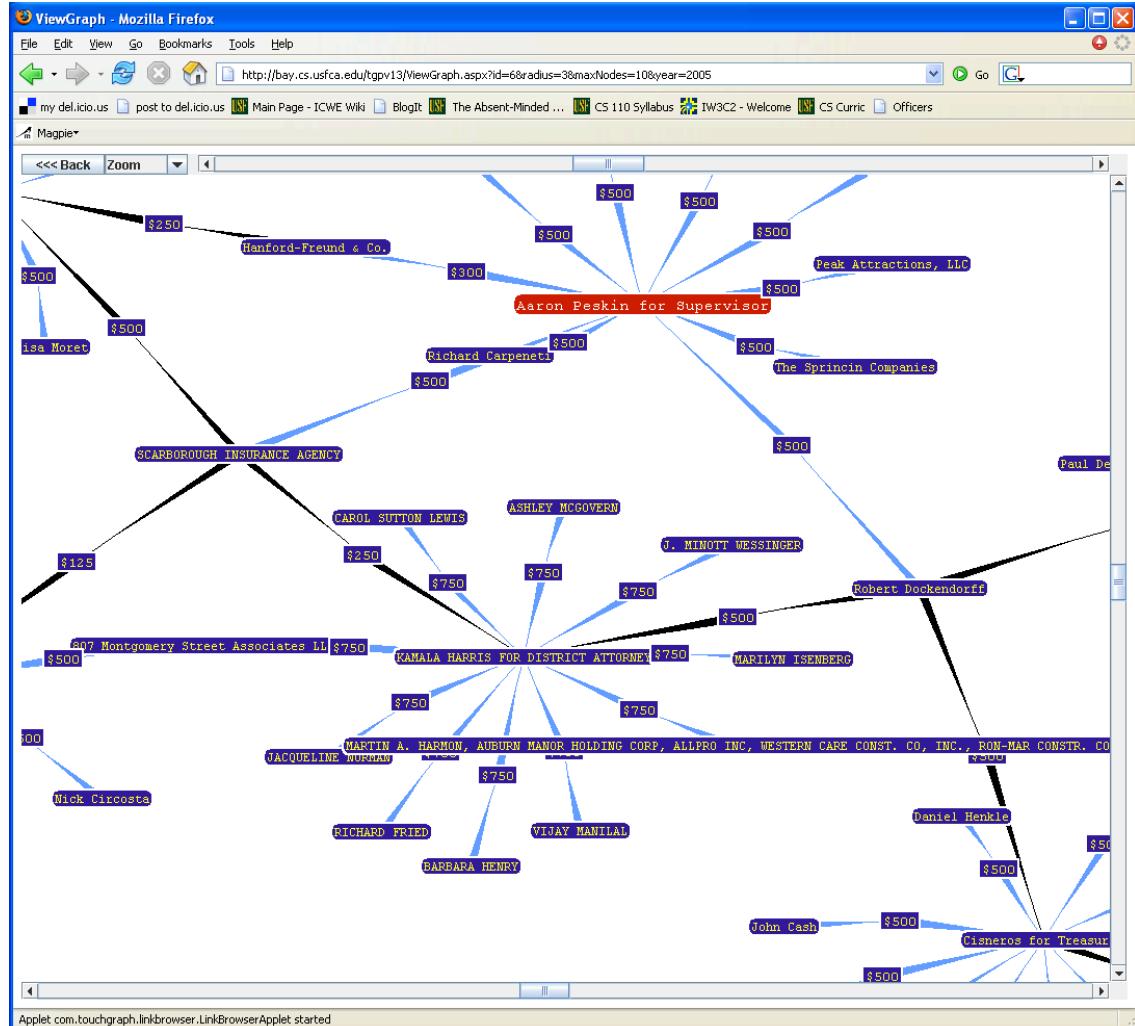


Figure 3. The Graph View of a Politician's Funding Relationships

So that the page can be updated without the intervention of a web developer or programmer, administrative pages were created that allow a non-technical staff member to enter the current office holders and important information about them. These forms, and the underlying database holding this basic information, is necessary because this data is not part of the campaign funding data at the Ethics Commission (it stores only funding transactions without information about the office of the entities).

Besides allowing the administrator to specify that, for instance, “Gavin Newsom” is the “Mayor”, the software also allows for the specification of an image and a homepage for each person, along with a list of the committees associated with the person. Though it easy for a human to know that “Gavin Newsom for Mayor” is a committee for “Gavin Newsom”, software cannot make such an assumption, so such human intervention is necessary.

The data displayed in Figure 1 combines the additional information provided by the administration forms and underlying database with funding data directly from the Ethics Commission site. The columns labeled ‘committees’ show only those committees that the administrator has specified as being associated with the officer. The funding summary data, and the data that can be seen if the link is followed, all comes from the campaign funding database.

Without this additional data, a user would have to either know the committees of a particular officer, or use a search page to scout for the associated committees. Furthermore, without the politician-committee relations being set explicitly in the database, it wouldn’t be possible for software to summarize data amongst the associated committees (whosfundingwhom.org doesn’t perform these calculations currently, but the database is designed to allow it).

3.2 Table-Based Navigation

When the user selects one of the entities from the landing page or the search screen, data concerning that entity is displayed in a tabular format, as in Figure 2.

In this example, the funding data for “Aaron Peskin for Supervisor” is displayed. Four tables are displayed. For a politician’s committee, the first two—“Top Contributors” and “Top Expenditure Recipients”—are generally non-empty. When a donor, independent committee, or vendor is displayed, other tables are more important (e.g., vendors generally have non-empty expenditures received).

The tables show the top ten associations. Each row in the table displays a *funding relationship*, which is a summary of all

transactions between two entities. Sometimes a relationship consists of just a single transaction, e.g., the first row shows a single transaction between “Aaron Peskin for Supervisor” and “Robert Dockendorf”. Sometimes two entities have shared more than one transaction. For instance, the second table in Figure 2 shows that there are 5 transactions between “Aaron Peskin for Supervisor” and “Stearns Consulting”.

Unlike many current systems, including the one at the San Francisco Ethics site, the user can navigate a money trail by selecting one of the entities listed. For instance, if “Stearns Consulting” is selected in the table of Figure 2, the user sees a listing of Stearns Consulting’s funding data, which shows that the firm also received expenditures from the San Francisco Women’s Political Action Committee.

3.3 Graph-View

While a table provides detailed information concerning a funding relationship, it shows funding trails at only one degree of separation. whosfundingwho.org also provides a graph view that shows entities and relationships with multiple degrees of separation (See Figure 3).

The graph shown in Figure 3 was invoked by selecting “Graph View” from the “Aaron Peskin for Supervisor” tabular view page. Both incoming and outgoing funding relationships are shown and color coded. The graph allows the user to view multiple-degree relationship chains, e.g., Peskin’s committee accepted a contribution from Robert Dockendorff who also contributed to Kamela Harris.

The graph view induces the biggest “aha!” reaction from those that have viewed it. At the unveiling of the software, Ethics commission staff and other interested parties immediately became engaged in discussions not about the software, but about the data. This slowed progress in terms of getting technical feedback but gave a great indication about how powerful data visualized in this way can be.

Eventually, such a graph could be designed so that relationships other than just money flow could be shown. For instance, the nodes for an individual could point to his or her associated committees, or the node for a committee might be connected to the lobbyists it has hired. The key to implementing such views lies in integrating data from various databases, as will be discussed in Section 5.1.

4. PLANNED REFINEMENTS

Plans for the next version of the “whosfundingwhom.com” software include the following refinements: 1) Eliminating unwanted aliases, 2) Providing ‘top ten’ lists and dynamic lists created by administrators.

4.1 Eliminating Unwanted Aliases

Certainly electronic filing improves the state of campaign information compared to the days of paper forms. Consider this statement:

“The old paper system made it difficult to track the flow of campaign cash . . . The candidates often did their best to keep the public in the dark. Former Governor Mario Cuomo’s reports regularly included handwritten entries, some illegible. [Governor George] Pataki filed printed reports, but used extremely small print and alphabetized his list of contributors for a time by first name.[4]

However, on-line filing in and of itself does not eliminate the problems presented in the above quote. If input is based primarily on typing free text into *unstructured* text fields, the problems with paper filings will persist in the electronic world, as filers can unwittingly or unwittingly refer to the same entity differently at different times.

For instance, in the raw data we download from the San Francisco Ethics Commission site, there are numerous examples whereby a single entity is referred to by various names. This is essentially an “input” problem caused by users typing entity names into unstructured text fields instead of choosing from data that is already in the database. With unstructured text, even an extra space or comma can lead to problems. For example, in San Francisco’s campaign finance data of 2004, “Haight Street Mortgage” appears as:

Haight Street Mortgage
Haight Street Mortgage Co Inc
Haight Street Mortgage Co., Inc.
Haight Street Mortgage, Inc.

A human can easily surmise that the text strings all refer to the same entity, but a computer program cannot make that assumption, so data will be displayed incorrectly.

It should not be assumed that such misinformation is due to wrong-doing. In this case, Haight Street Mortagage had no part in the problem-- various other entities filed “received payment” forms and typed in the mortgage company’s name in slightly different ways.

The root of the problem is that many on-line input forms are just paper forms directly transferred to the computer. With such PDF-like forms, the user can only enter text in boxes and cannot choose from existing entries. This is an example of a more general problem in human-computer interaction—programmers modeling the on-line world too closely to the paper world and not taking advantage of what is electronically possible.

Web applications, as opposed to PDF-forms, can connect directly to the live database and allow the user to choose from existing entities. Such applications can also notify the user, as her or she types, that an entity with a similar name exists. Such simple facilities can reduce the count of unwanted aliases significantly.

With no control over the input side of the equation, the plan for whosfundingwhom.org is to tackle the problem by performing alias processing on the raw data downloaded from the Ethics commission. Like the specification of associations between individuals and committees, such processing will be a joint effort between software and human—the computer can flag similar names and ask the administrator to make the final call on whether the similar names refer to the same entity. Our plan is to add this capability as part of the administration forms, so that Ethics Commission staff can continually monitor the data to remove unwanted aliases.

4.2 Additional Lists

Currently, the system allows a user to find a particular entity through the landing page of top officers, through a direct search, or through an association. The plan is to also add ‘top n’ lists which will provide pages for 1) the top n contributors over a time period, 2) the top n receivers of contributions, 3) the top n spenders, and 4) the top receivers of expenditures.

Whereas these lists are pre-defined, the system will also allow an administrator to create additional lists on-the-fly, just as the top officers page is created now. Such form-based list creation adds great flexibility to a site as users, not just programmers, can specify the collections of data that are displayed. Amazon was an early proponent of user lists, allowing users to create lists of books within their Listmania framework, and integrating those lists within their searches. Jeteye (www.jeteye.com) has applied the idea more generally to the web, providing a site whereby users can create arbitrary lists of web pages, notes, and images.

For whosfundingwhom.org, the administrator will be given the ability to create lists of entities, e.g., a list of “current candidates”, a list of “republicans” in the city, or a list of “political consultants”. The underlying motivation is that politics is dynamic in nature, so the system needs to be flexible enough to allow the data shown on the site to change dynamically. And, because of the high cost of software and the lack of funds available to most government organizations, a key is that those changes can be made without a software/web developer.

5. FUTURE OF POLITICAL E-IDENTITY

Developing software for visualizing campaign funding data is the first step in USF’s Transparency in Government Project. The long-term goal is to provide comprehensive information about the public record of politicians.

With the current state of the web, there is a lot of information but little organization. Journalists and other interested citizens can now forage for information at sites put out by the politicians themselves, at non-partisan sites that attempt to provide objective information on politicians, and finally to blogs and other participatory sites that provide a forum for public discourse.

Key data points of a politician’s public record include campaign finance information, information about the lobbyists or consultants hired by the politician, the politician’s voting record and stand on issues, the politician’s appointees and appointers, the organizations the politician has awarded government contracts to, and the politician’s employment records.

Much of this information exists on the web, but it is scattered. Compiling a politician’s public record takes a single investigator days, weeks, or even months to collect it from the various web pages on which it resides. And because most of the information is not structured—it is free text—software cannot perform such a task in an automated manner.

There are two strategies for tackling the problem: one is to induce organizations to publish information in a standardized XML format. If data were published in this manner, instead of as free text “reports” on web pages, then automated software could collect, process, and display it in various ways. Specifically, a site like whosfundingwhom.org could access data from various services to provide a more comprehensive view of a politician’s identity. The site might access, for example, voting records from the city elections commission, appointee records from the mayor’s office, and directors information from the Chamber of Commerce, in standardized XML forms, then display the data in graphs such as Figure 3. Such a strategy, based on standardized web services,

is an example of a general movement in information systems to a more *semantic web* [2].

The second strategy involves collaborative tools that can harness the time and effort of the many politically interested individuals now roaming the web. The idea here is that, until the day when all information sources publish data in a standardized XML format, there will be a need for *humans* to read web pages and extract the data that is pertinent to a politician’s public record. Whereas such a task is too enormous for a single individual or group, it could be possible with the combined efforts of the collective.

With this strategy, a site like whosfundingwhom.org would provide publicly accessible input forms for entering information about politicians and political organizations into the comprehensive system. Some of these forms now exist—forms for specifying who the current officers are, forms for specifying the committees of each politician, and forms for data like a politician’s home page and image. These forms could be opened to the public, and could be extended to allow for the entry of various types of data and relationships.

Such public collaboration has gained wide acclaim with the success of such efforts as wikipedia.org, an on-line encyclopedia that allows anyone to create and edit content. But can a collaborative strategy be used to collect and organize political information? And is it possible to harness the power of the many while also staying objective?

5.1 Semantic Web

Tim Berners-Lee, one of the founders of the web of today, is the leader in the movement towards the creation of a semantic web [2], one that consists of structured data and not just web pages. The semantic web is based on information systems providing web service access to their data, and providing input forms that enable the creation of structured data.

5.1.1 Web Services

Web services are distributed programs that allow client software to query another computer for particular data. Unlike normal web page requests, which return HTML, web service requests return data in a machine-readable XML format. In essence, web services provide the data without all the presentation formatting of HTML, from which software has a difficult time in extracting desired data.

There has been little progress towards providing web service access to campaign finance data. At the federal level, the FEC site (<http://www.fec.gov/disclosure.shtml>) does publish data in a structured format that can be read by computers, but it is in the form of a single downloadable file, not a web service, so the only way to use it is to download it in its entirety.

If the FEC site provided web service access to its information, a client program could request, for example, all the contributors of a particular politician. The results would be returned in a structured manner, without presentation formatting, so that the client program could process it, combine it with other data, and display it in any way it wants.

Such web services are the key to the interoperability problems involving databases “owned” by various sources and in various

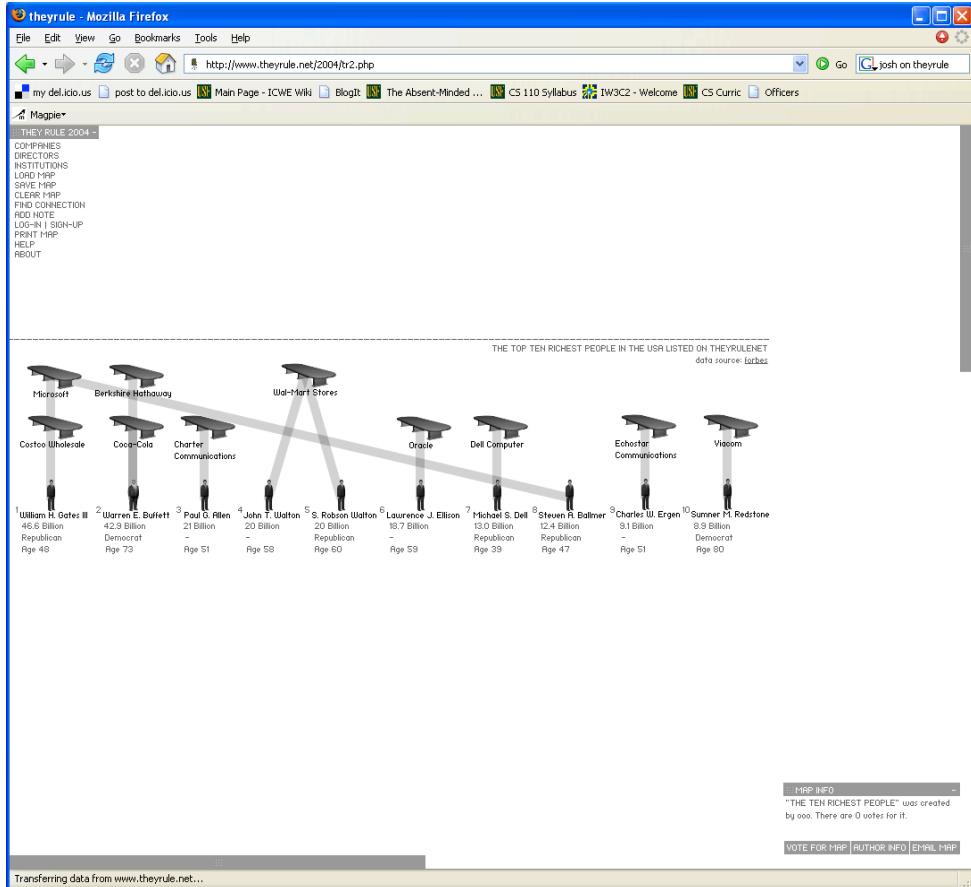


Figure 4. theyrule.net map displaying board member relationships between the top ten richest individuals.

formats. If interested parties can agree on a standard XML format for queries like “return all contributors of politician X”, then client programs can query different web services, e.g., the federal campaign finance service, the local campaign finance service, the lobbyist service, and a board of director service. Programs can then process the various data and display it in a comprehensive way. A user could then go to one site, enter a politician’s name to view all information about the politician, and be able to follow associative trails, e.g., view information about the politician’s lobbyists.

5.1.2 URI-based Input

Allowing a user to choose from existing entities, as discussed in section 4.1, can provide significant improvement in an information system, but that in and of itself does not eliminate the potential for ambiguity. Semantic web proponents also consider issues of identity through the use of Uniform Resource Identifiers (URI). URIs are like URLs in that they are globally unique identifiers, but unlike URLs they do not necessarily map to a file on some server.

What URIs can provide is help in uniquely identifying real world objects like people. Instead of referring to “David Wolber” in a web page to refer to the author of this paper, one could refer to “cs.usfca.edu/wolber” which is the unique uri for David Wolber, the USF professor.

Of course filers cannot be expected to know the URIs of individuals. What is needed is a global name service which client software can access, and which provides a mapping between basic

information (last name, first name, etc.) and a URI. When a filer enters a name, this service would be queried to return a list of matching known individuals. The user would then be allowed to choose from the list. Beneath the surface, the software would identify the individual using the URI from the service.

5.2 Public Participation

The previous section discussed one strategy towards building comprehensive political information sites: inducing information sites to publish their data in web service form, which would allow other software to access data from various places in order to provide comprehensive views of a politician’s public record. But such an “automated” strategy is dependent on many organizations agreeing to publish data in this way, then agreeing on a standard, and finally actually implementing the changes. Such a process may take years and it is doubtful that all pertinent information will be made available in XML form.

The key problem is that most data is in web page report form, which is understandable by humans but not software. A human can read it and understand it, and if given the time could research thousands of pages on the web and come up with a cohesive collection of data. For instance, Josh On researched the web pages of hundreds of companies to compile a comprehensive list of the board of directors in corporate America[7]. He then made this information publicly accessible at the website <http://www.theyrule.org> (See Figure 4).

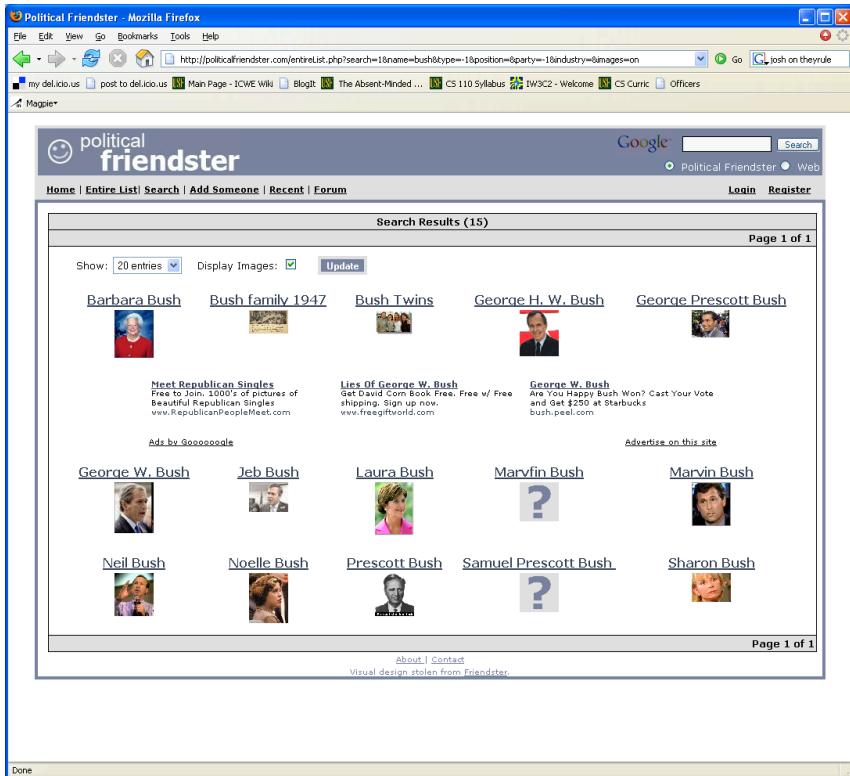


Figure 5. Political Friendster allows the public to enter people and relationships.

Unfortunately, humans don't have unlimited time to perform such work. [theyrule.net](#), for instance, has only 2004 data and is not kept current.

Can software perform such a task? Well, as discussed, software is best able to read structured data—data in pre-defined tables and fields. Natural language processing has progressed, but is not near the point in which a software program could crawl the web to discover the websites of all Fortune 500 websites and extract the board of directors information from them.

The semantic web solution, for [theyrule.net](#) would be to require corporations to file a list of their board of directors, and then have the government make that information available in a machine-readable XML format. Short of such automation, there is also the strategy of engaging the masses. There are thousands of intelligent people roaming the Internet along with emerging collaborative tools which can tap their enormous potential. Wikipedia ([www.wikipedia.org](#)), the on-line encyclopedia, is probably the most popular site that is edited by all that choose to contribute.

Another is [www.de.li.co.us.com](#), an on-line bookmarking system that allows users to categorize web pages and together create a *folksonomy* of popular web pages. A third example is Slashdot ([www.slashdot.com](#)) which allows users to submit and rate technology “stories”. The basic idea in all of these systems is to tap the power of the “technorati” on the Internet by letting everyone become not just consumers of information but publishers of information (*prosumers*).

The power of such collaboration has barely been tapped in the political world. [www.politicalfriendster.com](#) is one of the earliest political sites allowing the public at large to enter data. It allows any user to add a new person or organization to the database and add associations between entities, e.g., George H. Bush is senior

advisor to the Carlyle Group and a close friend of Prince Bandar Bin Abdul Aziz. Such relationships probably cannot be found in any existing database, only in news articles and documentaries. By entering such relationships into Political Friendster, a user brings them into the structured data world and potentially to any software that focuses on displaying political relationships (Political friendster currently does not publish its data in XML format, however, so one can only view the associations through that website).

The key issue with participatory sites is integrity of data and the potential for erroneous information to be input. Wikipedia has recently had some high profile problems, including one involving a man being accused of involvement in the Kennedy assassination [9], and another involving staff members of Senator Diane Feinstein altering reports concerning her husband's investments [5]. The stakes might even be higher for a site focused on politics.

But should public participation in data gathering be disallowed just because the data is so sensitive? Political blogs, after all, have even worse potential for erroneous information as users can enter any unstructured text.

The issue is to determine an appropriate level of site moderation and whether the moderation should be administered by a trained individual, the users themselves, or a combination thereof.

Slashdot, the technical news site, takes the latter approach. A team of individuals from the organization chooses which stories to “publish” from all those input by the public. Users then comment on stories. Because a story can have hundreds or even thousands of comments, a reputation mechanism is provided whereby users rate each other's comments. Those that have had highly rated comments in the past are given high reputations and their comments are listed more prominently.

Because users themselves help moderate the site, a significant burden is removed from the site maintainers—Slashdot would have to hire thousands of employees to keep up with the work. As non-partisan political information sites are often maintained by non-profit organizations, it is likely that funding will be scarce and public moderation will be a necessity. The reader is referred to [6] for a discussion of reputation and moderation schemes.

5.3 Next Steps

We plan to explore both the semantic web and public participation directions within the Transparency in Government project. The idea of opening the site to public participation has received less than enthusiastic support from both the McCarthy Center board of directors and the Ethics Commission, who both have advocated for limiting the site to objective data only. One possible compromise is allowing human-input but restricting it to a password-protected user-base consisting of trained administrators. Of course this solution severely restricts the amount of data that can be included since it doesn't harness the power of the masses. Other potential solutions include limiting the types of data that can be input (e.g., no free text comments, just associations) and providing extensive moderation to the site.

On the semantic web side, we are conducting a survey of existing XML-based initiatives in government, beginning with the list at XML.gov, and including inquiries into various local agencies. We also plan to implement a web service interface to our augmented campaign funding data, and explore the possibility of initiating the development of a standard protocol for such data.

6. SUMMARY

Early feedback of whosfundingwhom.org suggests that the software can significantly increase the ease of investigation. We see journalists and ethics commission staff as the key audience for the software, though those interested in identifying funding sources have also shown interest.

Besides the value of its end-product, the Transparency in Government Project at USF can also serve as a model in terms of university-city collaboration and service learning within the computer science curriculum. Many computer science students are well-skilled by the time they are in the upper classes. Instead of working on “academic” problems, they can apply their energies to real problems and help the community. The students who have participated in “whosfundingwho” have gained invaluable experience not often gained in a school setting—requirements gathering, working with people, giving presentations to groups including VIPs like Leo McCarthy and Art Agnos, human-computer interaction analysis, and exhaustive software testing. The students have also been exposed to the possibility of applying their craft in a public service setting, which is unfortunately all too uncommon in computer science education.

Though the project has practical goals, the potential for applied research is great given the relative infancy of both the semantic web and collaborative web sites. Semantic web research is in great need of problems that can really help someone: political data on the web can provide a great test bed for researching large-scale web service based data aggregation, as well as semantic analysis and inference techniques.

Collaborative tools like del.icio.us have just recently popularized the concept of mass annotation of the web, so the area is not yet well understood. Applying the idea to the political data on the web

should present and amplify many of the interesting issues, including privacy, techniques for fair and useful moderation, and issues of automated reputation measurement.

7. ACKNOWLEDGMENTS

This project has been funded by the City of San Francisco as well as the Leo. T. McCarthy Center at USF. Credit for the site goes to the following students who have contributed significantly to the “whosfundingwho.org” site and/or the lobbyist filing system we developed for the Ethics Commission:

Carmel Avnon
Cindy Zhao
Erick Chin
Pooja Garg
Vineet Agarwal
Uddhav Gupta
Monica Agarwal
Alex Lagor
Marc Greenberg

A special thanks also to Oliver Luby of the San Francisco Ethics commission, who knows more about campaign funding data and laws than any human alive. If a brain-transfer were possible, our software would be complete.

8. REFERENCES

- [1] Alexander, K., *Show Me the Money, Part 7: Emerging Digital Disclosure*.
http://www.rtndf2/publications/show_money7.html
- [2] Berners-Lee, T., Hendler, J., Lassila, O., *The Semantic Web*, Scientific American, May, 2001.
- [3] Institute for Politics, Democracy, and the Internet, *The Virtual Trail: Political Journalism on the Internet*,
http://www.pewtrusts.com/pdf/pp_online_journalist.pdf.
- [4] Holman, C. and Stern, R. *Access Delayed is Access Denied: Electronic Reporting of Campaign Finance Activity*. Public Integrity, Winter 2000.
- [5] Lochhead, Carolyn, *Staff Altered Online Entries on Feinstein, Blum*, San Francisco Chronicle,
<http://www.sfgate.com/cgi-bin/article.cgi?file=/c/a/2006/02/10/MNGSSH62CI1.DTL> Feb. 10, 2006.
- [6] Masum, H., and Yi-Cheng, Z. Manifesto for a Reputation Society, First Monday, Vol. 9, No. 7., July 5, 2004.
- [7] On, Josh, *From They Rule to We Rule: Art and Activism*,
http://www.aec.at/en/archives/festival_archive/festival_catalogs/festival_artikel.asp?iProjectID=11803.
- [8] Public Citizen: Protecting Health, Safety, and Democracy.
<http://www.citizen.org/congress/campaign/issues/disclosure/index.cfm>
- [9] Seigenthaler, John, *A False Wikipedia Biography*, USA Today,
http://www.usatoday.com/news/opinion/editorials/2005-11-29-wikipedia-edit_x.htm Nov. 29, 2005.

A Project to Assess Voting Technology and Ballot Design

Michael W. Traugott
Institute for Social Research
University of Michigan
Ann Arbor, MI 48109
734-763-4702

mtrau@umich.edu

Paul S. Herrnson
Ctr. for American Politics & Citizenship
University of Maryland
College Park, MD 20742
301-405-4123

pherrnson@capc.umd.edu

Richard G. Niemi
Department of Political Science
University of Rochester
Rochester, NY 14627-0146
585-275-5364

niemi@rochester.edu

ABSTRACT

In this paper, we describe the results of a study employing multiple designs and measurements to assess new voting technology and its impact on a variety of voter behaviors and attitudes. While we find the new devices improve generally create satisfaction with the experience and confidence that votes are recorded accurately, there are differences in assessments of different technologies. And issues of the “digital divide” are present.

Categories and Subject Descriptors

B.8.2. Hardware Performance Analysis and Design Aids
J.4. Computer Applications in the Social and Behavioral Sciences
K.4.1. Computers and Society: Public Policy Issues

General Terms

Performance, Design, Experimentation, Human Factors.

Keywords

Voting technology, usability, digital divide.

1. INTRODUCTION

This project looks at the impact of new voter technology on voting behavior, satisfaction with new voting technology, and confidence in the electoral system. The work is prompted by concerns that arose out of voters' experiences during the 2000 presidential election, especially in Florida, and continued through the 2004 election. The Help America Vote Act of 2002 mandated the abolition of certain kinds of voting machines and provided funds for the purchase of new equipment in accord with a set of voluntary guidelines for utilization but not usability.

This project adopts a multi-design data collection and a variety of associated analytical approaches to assess voters'

responses to the new voting technology. The research focuses on a generally accepted goal that the new technology should reduce “lost” or “residual” votes, but it goes further in looking at the ability of the new machines to capture “voter intent.” It also looks at satisfaction with the voting experience, and it is the first project to look at voting machine comparatively. One goal of the effort is to develop general purpose data collection and analysis procedures that will generalize to future elections.

2. RESEARCH DESIGNS

We assembled six different voting machines from five current manufacturers as well as a prototype that uses a “zoomable” device to highlight specific contests (offices or issues). The project has employed expert reviews, close observation of a few individuals voting on the machines in a usability lab, field data collection in malls and other locations like senior citizens’ homes, and the analysis of actual election returns in locales where the technology has changed and stayed the same as the basis for comparison. In addition to the analysis of data at the individual-level collected in the malls and the usability lab, we also developed techniques for ecological regressions applied to small-unit electoral data. Employing the equivalent of a “natural experiment” we analyzed voting behavior places that changed their technology between 2000 and 2004 and those that did not, in Florida and Michigan.

3. PROJECT ACCOMPLISHMENTS

All of our original data collection is completed. Most of the data are in usable form with some coding of the videotapes to be completed.

We have completed preliminary analyses of: 1) The expert evaluations of the 6 different machines; 2) Satisfaction with the voting experience from the field studies; 3) The relative accuracy of the machines; 4) The impact of changing technology on residual votes in two case study states; 5) Analysis of write-in procedures; and 6) Usability factors associated with voting on some of the machines

In addition to attending professional conferences in a number of disciplines, we have participated in special meetings held at the American Association for the Advancement of Science, National Academies of Science, National Association of State Election Directors, and the National Institute for Standards and Technology dealing with voting technology.

4. FINDINGS

For our studies, we constructed a “simulated” election consisting of 22 partisan and nonpartisan offices and referenda. Participants used a Voter’s Pamphlet to review their options and to select their choices. Beyond this, participants were also asked to do two things that voters commonly do: switch from one candidate to another for one office and write in the name of a candidate for another office. This pamphlet became our indicator of their “intent.” Participants carried this booklet with them as they voted on 6 different machines to which they were randomly assigned.

In the usability lab studies, voters could make comparative judgments about the relative ease and convenience of using the 6 machines [1]. By coding videotapes for duration of actions as well as distinctive paths of movements, we developed a sense of what might produce satisfaction or dissatisfaction with a particular device.

In field tests, voters seemed to navigate the new technology well even with limited training and a simulated election. They expressed greater confidence in the DRE machines to record their votes accurately, especially the Diebold and “zoomable” devices. [2] There was a suggestion of access and satisfaction issues associated with the “digital divide.” When we assessed the accuracy of their voting [3, 4], overall about 97% of their actions were on target, although 20% of them made at least one “error” when voting. The number of errors increased as the task became more complex (changing votes, selecting two candidates for a single office, making a change after voting for a given candidate, and using a straight-party ballot). Issues of the “digital divide” were very apparent in these analyses.

An analysis of election returns in 2000 and 2004 in Florida and Michigan showed that the adoption of new voting technology succeeded admirably [5]. In both states, every change from old to new technology was accompanied by a decline in the rate of residual votes; the declines in places that introduced new technology were significantly greater than in areas that did not change. Where the residual vote was especially high, as in some places in Florida that used punch cards in 2000, the decline was extraordinary—as high as a 96% reduction. Declines in residual vote rates also occurred even in areas that did not change their voting equipment. Attention given to voting procedures evidently led to better ballot design, improved voter education, more alert voters, or all three.

Taken together, the results show that usability tests should be conducted on all (new and existing) voting systems in order to improve them. More attention should be given to ballot design; the layout of the ballot, the use of offices or parties as the means for organizing the ballot, the use of straight-party features, and the number of choices to be made can influence voter accuracy. Training programs should be designed to help voters understand the process of voting. Recognizing the pace of technological change, along with the mobility of the population, training should

ideally teach people underlying concepts (such as straight-ticket voting and multiple votes for the same office) along with the skills needed to utilize voting systems adopted in their locality or state. Training should be made available at polling places on Election Day. Election officials should be trained and employed in anticipation of the problems certain groups of voters, such as the elderly or minorities, can be expected to encounter when voting. Election judges and voting systems should be deployed in greater numbers in precincts where large numbers of these individuals are likely to vote so that those who are the most likely to need help or who take longer to vote can be accommodated.

5. ACKNOWLEDGMENTS

We gratefully acknowledge the support of the National Science Foundation for our research (Grant Number 0306698) and the Carnegie Corporation for a grant to support outreach (Grant Number D05008). are the sole responsibility of the authors.

6. REFERENCES

- [1] Conrad, F. G., Peytcheva, E., Traugott, M. W., Hanmer, M. J., Herrnson, P. S., Bederson, B. B., Niemi, R. G. Voter Intent, Voting Technology and Measurement Error. Presentation at the annual meeting of the American Association of Public Opinion Research. Miami Beach FL (May 2005).
- [2] Herrnson, P. S., Niemi, R. G., Hanmer, M. J., Bederson, B. B., Conrad, F. G., and Traugott, M. W. The Not-So-Simple Act of Voting: An Examination of Voter Errors with Electronic Voting. Paper presented at the annual conference of the Southern Political Science Association, Atlanta GA (January 2006).
- [3] Herrnson, P. S., Niemi, R. G., Hanmer, M. J., Francia, P.L., Bedersen, B. B., Conrad, F., Traugott, M. W. The Promise and Pitfalls of Electronic Voting: Results from a Usability Field Test. Paper presented at the annual conference of the Midwest Political Science Association, Chicago IL (April 2005).
- [4] Herrnson, P. S., Niemi, R. G., Hanmer, M. J., Francia, P.L., Bedersen, B. B., Conrad, F., Traugott, M. W. How Voters React to Electronic Voting Systems: Results from a Usability Field Test. Paper presented at the annual conference of the Midwest Political Science Association, Chicago IL (April 2005).
- [5] Traugott, M. W., Hanmer, M. J., Park, W. H., Niemi, R. G., Herrnson, P. S., Bederson, B. B., Conrad, F. G. Losing Fewer Votes: The Impact of Changing Voting Systems on Residual Votes. Paper presented at the annual conference of the American Political Science Association, Washington DC (September 2005).

E-voting in the 2005 Local Elections in Estonia and the broader impact for future e-voting projects

Alexander H. Trechsel
European University Institute
Via dei Roccettini 9
San Domenico di Fiesole (FI), Italy
+39 055 4685 442
Alexander.Trechsel@iue.it

Fabian Breuer
European University Institute
Via dei Roccettini 9
San Domenico di Fiesole (FI), Italy
+39 340 7364653
Fabian.Breuer@iue.it

ABSTRACT

In this project, we analyze the introduction of voting by internet (e-voting) in political elections. In particular, we focus on the Estonian municipal elections held on 16 October 2005, where the possibility of casting a vote via the internet was newly introduced. This introduction of e-voting represented a true *world-première*: even though internet voting has so far been used in consultative decision making processes around the globe, in private elections and in a number of formally binding referendums, the local elections in Estonia were the first time that an electorate of an entire country could cast its vote over the Internet in a public election. On behalf of the Council of Europe, and in close collaboration with the Estonian authorities, our research team academically accompanied this landmark event in e-democracy. Completing the process-tracing of the internet voting implementation, we conducted a survey among over 900 citizens, allowing us to conduct an in-depth study on voting behaviour in e-elections.. Our research shows what voting channels have been used by what type of citizens and contains a fine-grained analysis of participation patterns and political behaviour of the citizens in these elections. Finally, the data and findings of our research allows us to make selective comparisons with other studies in the field of e-democracy, which we conducted in several elections and referendums in Switzerland.

Categories and Subject Descriptors

J.1 ADMINISTRATIVE DATA PROCESSING GOVERNMENT

General Terms

Documentation, Human Factors, Theory

Keywords

E-voting, e-democracy, participation, elections, voting channels.

1. MOTIVATION AND BACKGROUND

The issue of e-voting has increasingly become a controversial topic among both political commentators and practitioners. There is a growing scientific and public debate on the issue, which in some cases arouses great passions. Overall, e-voting is considered to offer citizens new voting tools in order to slow down the

perceived erosion of participation rates in many Western democracies. In addition, e-voting is seen as part of an overall governmental reshaping process of its own administrative processes and outputs.

E-voting can be placed within its wider theoretical context to the concept of e-democracy. We consider e-democracy to consist of all electronic means of communication that enable/empower citizens in their efforts to hold rulers/politicians accountable for their actions in the public realm. Depending on the aspect of democracy being promoted, e-democracy can employ different techniques. In the case of e-voting, the aim is to enhance the direct involvement and participation of citizens in the democratic process.

Even though there exists an emerging theoretical debate on the issue of e-democracy and e-voting, empirical research in the field is quite limited so far. Given the fact that particularly the tool of e-voting was only applied very rarely up till now, this is not surprising. As already mentioned, the local elections held in Estonia in October 2005 present an exception in this regard: the elections were the first time that an electorate of an entire country could cast its vote over the internet in a public election.

In general, Estonia - which is often referred to as "e-stonia" - is considered to be a leading country when it comes to the use of ICT-technologies and the internet in the private as well as in the public sector. The majority of the population uses the internet, all Estonian schools are connected and 750 public internet access points exist. Income tax declarations can be made electronically and online, expenditures made by the government can be followed on the internet in real-time and cabinet meetings have been changed to paperless sessions using a web-based document system.

The local elections gained considerable international attention due to their "first-time-ever" character as well as to their wider significance for designing secure remote voting systems. There were observers from approximately 30 countries who wished to follow the functioning of the new voting channel by internet. The possibility of online voting is under scrutiny in many countries and has been promoted as a quick and cheap way of collecting

ballots. Apart from that, supporters of e-voting argue that e-voting could bolster democracy by increasing participation in elections. Amongst others, the political authorities of the Kingdom of Bahrain announced that they are interested in using parts of the Estonian e-voting system for their parliamentary elections in 2007.

This *world-première* obviously is of crucial interest for the scientific and policy communities in theoretical and practical regards. On behalf of the Council of Europe we conducted a study on the elections and we were able to academically accompany elections. Among other things, we conducted a telephone survey among the Estonian electorate to research the use and the attitudes towards the newly introduced voting channel by internet.

2. WORK TO DATE

The broad study we conducted on the Estonian local elections in December 2005 delivered rich and interesting data and results. We conducted a classical telephone survey, which allowed us to perform a significant and detailed analysis of voting channels and voting behaviour of the Estonian electorate. The specifically designed telephone survey was conducted among 939 Estonian voters who had the right to cast their ballot in the elections of 16 October (the sample consisted of 315 e-voters, 319 ‘traditional’ voters and 305 non-voters).

By analysing the participation patterns and the political behaviour of the citizens we were able to shed further light on a number of relevant socio-political questions concerning this new mode of political participation. The results of the survey and the research were published in a report for the Council of Europe, which uses it in a wider project on the development of ICT-technologies and their role for democracies and good governance.

Next to the findings on the elections in Estonia, we started to carry out selective comparisons with other survey data from online-surveys in Estonia as well as from surveys in other contexts. In order to allow for comparative insights, both the questionnaire and the sample of the Estonian survey were built on previously undertaken surveys, particularly those in the Canton of Geneva, Switzerland.

In January 2006 we had the possibility to present our preliminary findings to interested authorities in the Kingdom of Bahrain in the framework of the E-voting Forum in Bahrain. In addition, we presented the findings as well as related policy recommendations to the Estonian National Electoral Committee in Tallinn. Overall, various national and regional authorities are interested in the research findings in order to use them in the development of own e-voting projects.

3. PRESENTATIONS

We have presented this work at the following venues:

World Summit on the Information Society in Tunis (Tunisia)

Bahrain E-voting Forum (Kingdom of Bahrain)

Estonian National Electoral Committee in Tallinn (Estonia)

University of Tallinn (Estonia)

University of Geneva (Switzerland)

Various workshops and seminars at the European University Institute, Florence (Italy)

ACKNOWLEDGMENTS

Primarily, we would like to thank the Council of Europe for sponsoring and enabling the research on the local elections in Estonia. Furthermore, we thank the research team in Estonia, and in particular Ms. Liia Hänni of the “e-Governance Academy” for their invaluable help and assistance with the data gathering. We also thank Fernando Mendez and Uwe Serdült of the “e-Democracy Centre” for their feedback on the survey questionnaire.

4. REFERENCES

- Gritzalis, Dimitris. *Secure Electronic Voting (Advances in Information Security)*, Kluwer, 2003.
- Trechsel, Alexander and Mendez, Fernando. *The European Union and e-voting : addressing the European Parliament's internet voting challenge*, Routledge, 2005.
- Trechsel, Alexander H./Kies, Raphael/Mendez, Fernando/Schmitter, Phillippe C. *Evaluation of the use of new technologies in order to facilitate democracy in Europe. E-democratizing the parliaments and parties of Europe*. Report for the European Parliament Scientific and Technological Options Assessment Series (STOA Report), 2004.
- The homepages of the E-Democracy Center and the e-Governance Academy are at: <http://edc.unige.ch/> and <http://www.ega.ee/>
- Information on e-voting in Estonia is on the official homepage of the Estonian National Electoral Committee at: <http://www.vvk.ee/engindex.html> and <http://www.vvk.ee/elektr/docs/Yldkirjeldus-eng.pdf>
- For the activities of the Council of Europe in the project “Good Governance in the Information Society” see http://www.coe.int/t/e/integrated_projects/democracy/02_Activities/.

SESSION 2A

INTEGRATED JUSTICE

Moderator

Hans J. (Jochen) Scholl, University of Washington, USA

Titles and Authors

Research Issues Related to Exchanging Information from Heterogeneous Data Sources
Swigger, Kathleen; Brazile, Robert

Integrated Criminal Justice: ARJIS Case Study
Sawyer, Steve; Tyworth, Michael

COPLINK Center: Social Network Analysis and Identity Deception Detection for Law Enforcement and Homeland Security Intelligence and Security Informatics: A Crime Data Mining Approach to Developing Border Safe Research
Chen, Hsinchun; Atabakhsh, Homa; Wang, Alan G.; Kaza, Siddharth; Tseng, Lu Chunju; Wang, Yuan; Joshi, Shailesh; Petersen, Tim; Violette, Chuck

Research Issues Related to Exchanging Information from Heterogeneous Data Sources

Kathleen Swigger

University of North Texas
Box 310440
Denton, Texas
Kathy@cs.unt.edu

Robert Brazile

University of North Texas
Box 310440
Denton, Texas
Brazile@cs.unt.edu

ABSTRACT

This paper highlights a project designed to help FEMA's Region VI address their information needs as well as extend the current technology for integrating heterogeneous databases. The successful completion of this project should lead to new and innovative ways to integrate multiple autonomous data sources. On the experimental side, the project is examining (1) architectural solutions to constructing mediated databases, and (2) new methods for supporting collaborative database activities. More specifically, the project is demonstrating the feasibility of representing the mappings between the virtual global database and the various data sources as XQuery queries [4]. We have chosen to use XQuery because of its efficiencies in processing the data, its ability to describe the data mappings, and its potential for improving standardization. The project's second task will address the collaborative interface problem. We are extending our current interface to support database activities such as viewing reports, sharing information, and editing data.

Categories and Subject Descriptors

H.2.5 [Heterogeneous Databases], H.2.5.2 [Information Interfaces].

General Terms

Management, Design, Human Factors.

Keywords

Heterogeneous databases, metadata, mediator systems collaborative software, groupware.

1. INTRODUCTION

As most everyone now knows, the Federal Emergency Management Agency (FEMA) is tasked with responding to, planning for, recovering from and mitigating against disasters. As part of this mission, FEMA is responsible for monitoring all nuclear power plants throughout the country. Each nuclear power plant in the United States has been required to have both an onsite and offsite emergency response plan. Onsite emergency response plans are approved by the Nuclear Regulatory Commission (NRC), while offsite plans (which are closely coordinated with the utility's onsite emergency response plan) are evaluated by the Federal Emergency Management Agency (FEMA) and provided to the NRC, who must consider the FEMA findings when issuing or maintaining a license. Federal, State and local officials work together to develop site-specific emergency response plans for nuclear power plant accidents. These plans are then tested through exercises that include protective actions for schools and nursing

homes. The plans also delineate evacuation routes, reception centers for those seeking radiological monitoring and location of congregate care centers for temporary lodging.

Until recently, FEMA's records about nuclear site visits and extended exercises were maintained largely by hand. Although some of the bigger regions developed small database systems to manage local information, most regional administrators assembled emergency plans and exercises by consulting previous excel spreadsheets and word documents. Faced with the mounting data, FEMA agents began to explore a number of low-cost solutions to their data management problem. As part of this exploration, Region VI officials asked researchers at the University of North Texas to become involved in the project.

Therefore, the PIs began a research project that is supporting FEMA with their data integration efforts by helping them manage their nuclear regulatory processes. We have accomplished this task by developing a low-cost database solution that can assist and automate the data integration and report generation processes. Our pilot project has developed a prototype architecture and automated tools for integrating data gathered from nuclear sites throughout FEMA's Region VI. By the end of the project, this data will be merged with other Regions' data, which is now distributed throughout the country. The long-term goal of the project is meet the immediate needs of Fema, which is to provide nuclear regulatory oversite for its citizenry, but also to find more effective and efficient ways of integrating data for other FEMA functions.

2. RESEARCH ISSUES

2.1 Metadata Representations

Central to our work has been the development of a repository for the data (and knowledge). A data repository needs to be extensible and adhere to an open standard, thus providing the ability to receive data from multiple sources in a seamless way. XML meta-data has been developed to incorporate all of the necessary information for storing the volumes of data. Work on developing mediator databases has provided the basic framework for the knowledge repository that is being used to store the FEMA data [4]. However, one of the major challenges for mediator databases is the problem of generating the mappings in the metadata repository. Developing the necessary logic to transform the expressions can become very complicated. We chose to use XQuery because we believe it provides a more natural solution to this problem. XQuery queries can be written in XML format using XQueryX syntax to facilitate parsing and processing. XQuery is an expression language that can be used to extract the source data, transform it, and create new structures in an easily

understandable way. Thus, XQuery queries can be used to represent the complex extractions, transformations, and reorganizations required in the metadata repository. Using XQuery queries as metadata means that the intelligence for transforming the user query into atomic queries for individual data sources resides in the metadata repository and not in the mediator logic, which makes the mediator easier to program and understand. Since the transformed XQuery queries are generated by the mappings between schemas, they are easier to maintain as well [3, 4].

Since we are using XQuery to represent the mappings, we have extended the Marian and Siméon projection path algorithm to process those queries. In previous work, Marian and Siméon developed a projection algorithm for XML documents to enable memory-based XQuery engines to process larger documents [2]. However, this algorithm did not include the constructed elements, which play a key role in our particular mediation solution. Thus, we extended the Marian and Siméon projection to constructed elements and applied it to the data integration process. Our particular algorithm addresses the constructed elements and excludes some paths that do not contribute to the query. This extension to the projection algorithm enables memory-based XQuery engines to process even larger documents than Marian and Siméon's algorithm whenever queries involve project operations on constructed elements. This extension allows us to identify the relevant data needed to answer a query from the data sources, and to design a decomposition algorithm to rewrite a user query on mediated schema to queries on local schemas. In extending this algorithm, we have established static inference rules for our extended projection and proved its correctness.

2.2 Collaborative Interfaces

Each year, FEMA conducts exercises in which they verify that a nuclear site is in compliance of the federal regulatory laws. These exercises involve large groups of people that are assigned the tasks of monitoring a critical area such as hospital preparedness, proper evacuation routes, etc. This phase of the project requires team members to share their findings, write joint reports, and record their experiences. Thus, one of the outcomes of this project is the creation of a shared database environment that allows users to access data, search shared results, visualize vast amounts of data, and collaborate with one another over the Internet.

Much has already been accomplished in the way of developing synchronous applications that run over the Internet. For example, multiple users located in different cities can now edit the same document simultaneously; have real-time team meetings; engage in training programs; and play games with one another. For the past eight years, the PIs have been involved in research concerning support for distributed teams. One of the results of this effort is the development of a special International Collaborative Environment (ICE) [1], which allows groups to collaborate, both synchronously and asynchronously, over the Internet [6]. The advantages of ICE over other collaborative software tools such as NetMeeting or instant messengers is that it provides a flexible framework for creating new tools that can be customized to new users' specifications.

Thus, one of the outcomes of this research is the creation of a database system that allows FEMA officials to collaborate with one another. Since the International Collaborative Environment (ICE) [6] is a system that was implemented in Java as a collection

of components, we are modifying it to support the writing of the group reports. The system currently supports chat, drawing, whiteboard, file sharing, editing, browsing, desktop sharing, and emailing tools. More importantly, the system provides a flexible framework for creating new tools that can be customized for different users and groups. We have now extended ICE's components to include tools for displaying different types of data and their results.

In order to gage the ‘usability’ of these interfaces, the designs are being made subject to usability evaluations [5]. The assessment of the interfaces consists of two aspects: System and usability evaluation. The system evaluation will center on assessing the system’s ability to request and receive the necessary information. The usability evaluation centers on the effectiveness of our interfaces in making a difference in FEMAs ability to collaborate with different members of the exercise team.

3. CONCLUSION

The proposed research directly benefits FEMA’s Region VI by helping them capture and record information on nuclear sites. By the end of the proposed research, we will make available to FEMA and the research community: (1) a prototype architecture and a set of specifications that support data integration of FEMA’s nuclear site information among all the regions; (2) Special collaborative interfaces that can be used to query integrated databases more effectively. The research will also facilitate information exchanges among the different regions throughout the country with the intention of enhancing these collaborations. The infrastructure and tools acquired (or developed) during the project will also be used to design and develop other integrated information efforts in FEMA such as emergency management, disaster recovery, etc.

4. ACKNOWLEDGEMENTS

This work is being supported by NSF’s Digital Government grant (#0506024).

5. REFERENCES

- [1] Brazile, R., Swigger, K., Harrington, B., Harrington, B. and Peng, X. The international collaborative environment, *CAINE*, San Diego, October 2002.
- [2] Marian, A. and Siméon, J. Projecting XML documents, *VLDB 2003*, 2003.
- [3] Peng, X., Brazile, R. and Swigger, K. Using XQuery to describe mappings from global schemas to local data sources, *Information Reuse and Integration, 2004, IRI 2004*, , 97, 8-10 Nov. 2004.
- [4] Peng, X., Brazile, R., and Swigger, K. Extending XML document projection for data integration, *Information Reuse and Integration, Conf, 2005. IRI -2005 IEEE International Conference on* , 138- 143, Aug. 15-17, 2005.
- [5] Preece, J., Rogers, Y. and H. Sharp, *Interaction Design: Beyond Human-Computer Interaction*, John Wiley & Sons, New York, 2002.
- [6] Swigger, K., Brazile, R. and Monticino, M. Effects of culture on computer-supported collaborations, *Journal of Computer Human Interface*, 60, 3, 365-380, 2004.

Integrated Criminal Justice: ARJIS Case Study

Steve Sawyer

Michael Tyworth

School of Information Science and Technology

The Pennsylvania State University
University Park, PA 16802 USA
011-814-865-4450
sawyer@ist.psu.edu

School of Information Science and
Technology

The Pennsylvania State University
University Park, PA 16802 USA
011-814-865-4450
mtyworth@ist.psu.edu

ABSTRACT

In this paper we provide project highlights of our ongoing case study of an integrated criminal justice system (San Diego, California's Automated Regional Justice Information System or ARJIS). We develop this case to be used in a comparative analysis of other, similar, systems. Our focus is on better understanding and theorizing on complex web of relationships among work, the structure and governance of social institutions, and technological architectures. Our intent is to further principles of socio-technical design regarding computerization (an aspect of social informatics). Our work to date leads us to a set of concepts we are calling organic development. Organic development reflects a strategic use of top-down and bottom-up design principles, demands strategic leadership, and open design principles.

Keywords

Integrated Criminal Justice systems, institutions, architectures, organic development, systems design, case study, social informatics.

1. INTRODUCTION

We are conducting a case study of one integrated criminal justice system (the San Diego, CA area's Automated Regional Justice Information System or ARJIS (see www.arjis.org). We are using the data from this case study to compare with case studies we've completed on similar systems here in the United States and United Kingdom.

The other U.S. systems include the Washington, D.C. area's Capitol Wireless Integrated Network (CAPWIN, see www.capwin.org) and Pennsylvania's Justice Network (JNET, see www.jnet.state.pa.us). We have also been developing a case study of the United Kingdom's activities surrounding Airwave (their national wireless access to their nationalized policing, criminal justice and homeland systems, see www.pito.ac.uk).

The intent of this work is to theorize on how best to design complex inter-organizational systems. The context of policing, public safety, emergency response and homeland security provides both a topical domain and a conceptually rich space. The institutional complexity of this domain (due in part to the federal (and, thus, federated) structures of the U.S. government and due in part to the histories among many of the participating organizations) is a test of institutional force. The broad range of technological innovations, infrastructures, and trajectories mirrors the institutional complexity. The local and national desire to develop more information sharing, better collaboration, and interoperability provides impetus for change.

Our theorizing is focused on better understanding and predicting the outcomes of system design and usage activities on complex web of relationships among work, the structure and governance of social institutions, and technological architectures. Our intent is to further principles of socio-technical design regarding computerization (a central focus of social informatics).

2. TYPE OF COLLABORATION

With the help of the ARJIS leadership, we have been able to do several rounds of field work in the San Diego, CA area. We have also been able to continue our collaborative activities with JNET, CAPWIN, and PITO organizations. In addition, the various leaders and members of these organizations have, in turn, introduced us to other organizations, associations and contacts, and we are pursuing additional work with several.

To date, we have been unsuccessful engaging the Department of Homeland Security in this work, or gaining the attention of the National Institute of Justice/Department of Justice.

3. SCIENTIFIC RESEARCH OBJECTIVES

3.1 Accomplishments

To date we have completed all primary data collection on the ARJIS case. This complements the secondary data collection, which we continue to do. We have now begun full analyses of the ARJIS case data.

We have developed three interim findings: (1) ARJIS employs a Joint Powers Agreement as the basis for the organization. This places decision-making authority in a board of directors comprised of individuals from each member agency. (2) Instead of attempting to replace legacy systems, ARJIS has decided to

incorporate their legacy systems into the suite of applications while developing new applications using modern technologies that will eventually replace the legacy systems. (3) ARJIS' use is embedded into law enforcement work.

We are now beginning the comparative analysis across the four case studies.

3.2 Management Structure for Project

Steve Sawyer is working as principal investigator. Mr. Michael (Mike) Tyworth (a second year doctoral student at Penn State's School of Information Sciences and Technology) is participating as a research assistant. Mike has done much of the primary data collection.

We are working with organizational leaders and members from ARJIS and in contact with others at JNET, CAPWIN and PITO.

3.3 Collaboration examples

Ms. Pam Scanlon, director of ARJIS, has used some of our work in her briefings for ARJIS political and operational leadership. Mr. George Ake (late of CAPWIN) has supported our field work, participated at the D.GO 2005 conference, and drawn on our work in his briefings to sponsors and participants.

4. BROAD IMPACT

The *technical merit* of this work will be realized through three contributions:

1. Findings and generalizable principles regarding appropriate governance structures, operational plans and work arrangements to better support integrated criminal justice activities.
2. Design criteria to support development of future (and evolution of current) information systems to better support integrated criminal justice.
3. Findings and guidance regarding policy issues relative to information sharing; building and using integrated criminal justice systems and; and their governance.

The *broader social impacts* of this work will be reflected in two contributions:

1. The design and operational guidance will lead to improved systems and increased usage, leading to safer communities, greater homeland security, more pro-active policing and improved e-government.
2. The design and operational guidance will also provide insights into improving inter-agency and cross-level work and governance, and important aspect of improved government.

5. CHALLENGES AND OPPORTUNITIES

We see two challenges ahead relative to achieving the proposed technical goals:

1. Developing general principles from the comparative analysis.
2. Drawing effective guidance for developers and policy makers, developing the argument and presenting the material in ways (an

in venues) to reach these too-often disparate audiences.

To achieve the broader social impacts of this work we must: (1) engage the professional community by (2) developing depth of findings (needed to provide compelling vision for engaging).

6. RESEARCH VALUE OF WORKING IN DIGITAL GOVERNMENT DOMAIN

We are engaging digital government research for several reasons:

1. Public sector organizations and government can benefit from better uses of information and communications technologies.
2. Public sector organization and government have different needs and pressures than do private sector organizations. This includes different institutional structures and governance, specific and service-oriented missions, more complex (and required) inter-organizational interactions, differently structured budgeting, political pressures, and issues with technological infrastructure age, standards, funding and skill base.
3. A focus on information sharing that is both demanded by law and required for proper operations.

7. RECOMMENDATIONS FOR IMPROVING THE NSF DIGITAL GOVERNMENT PROGRAM

We are delighted to be a part of this program and can offer only simple and obvious suggestions:

1. Increase the amount of money available (perhaps via securing matching funding from federal and state agencies as partners)
2. Connect directly to Federal and State e-government and digital government initiatives (such as criminal justice/homeland security, enterprise architecture, electronic voting...).
3. Provide support for appropriate international collaboration (with, say criminal justice/homeland security, relief efforts, and non-governmental organization coordination (e.g., the UN)).

8. ACKNOWLEDGEMENTS

We thank the officers, professional staff, and other stakeholders in the ARJIS activity for allowing us to engage them during the field work. Special thanks to Ms. Pam Scanlon her assistance.

9. REFERENCES

1. Sawyer, S., Reagor, S., Tyworth, M. and Thomas, J. (March, 2005) "From Response to Foresight: Managing Knowledge and Integrated Criminal Justice," in Newell, S. and Galliers R. (eds.) *Proceedings of the 2005 Organizational Learning and Knowledge Capabilities Conference*, Cambridge, MA, 17-19 March, 2005.
2. Tyworth, M. and Sawyer, S. "Organic Development: A Top-Down and Bottom-Up Approach to Design of Public Sector Information Systems," (For inclusion in the D.Go 2006 Conference, May 22-24, San Diego, CA).

COPLINK Center: Social Network Analysis and Identity Deception Detection for Law Enforcement and Homeland Security

Intelligence and Security Informatics: A Crime Data Mining Approach to Developing Border Safe Research

Hsinchun Chen, Homa Atabakhsh, Alan G. Wang, Siddharth Kaza, Lu Chunju Tseng, Yuan Wang, Shailesh Joshi, Tim Petersen*, Chuck Violette*

University of Arizona, Department of Management Information Systems,
Artificial Intelligence Lab. and *Tucson Police Department

{first-name}@eller.arizona.edu

{first-name.last-name}@tucsonaz.gov

ABSTRACT

In this paper, we describe the highlights of the COPLINK Center for law enforcement and homeland security project. Two new components of the project are described, namely, identity resolution and mutual information.

Categories and Subject Descriptors

H.2.8 [Database Management]: Database Applications – *data mining*.

G.3 [Probability and Statistics]: *Probabilistic algorithms*

General Terms

Algorithms, Security

This research was supported in part by the NSF Digital Government (DG) program: "COPLINK Center: Information and Knowledge Management for Law Enforcement" #9983304, NSF Knowledge Discovery and Dissemination (KDD) program: "COPLINK Border Safe Research and Testbed" #9983304, NSF Information Technology Research (ITR) program: "COPLINK Center for Intelligence and Security Informatics Research - A Crime Data Mining Approach to Developing Border Safe Research" #0326348, and Department of Homeland Security (DHS) through the "BorderSafe" initiative #2030002.

Keywords

Identity Matching, Mutual Information, Data Mining, Naïve Bayes model, Semi-supervised Learning, Intelligence and Security Informatics, Law Enforcement, Border Safety, Homeland Security

1. INTRODUCTION

The COPLINK center [1] is a partnership between the University of Arizona's Artificial Intelligence Lab. and the Tucson Police Department (TPD). Other law enforcement partners include the Tucson Customs and Border Protection (CBP), Pima County Sheriff's Department, and San Diego Automated Regional Justice Information Systems (ARJIS). The objectives of the COPLINK center include the development of tools for the cross-jurisdictional collaboration, sharing, management and visualization of law enforcement data while keeping the data private and secure.

2. COPLINK CENTER: Challenges and Accomplishments

2.1 Cross-Jurisdictional Collaboration

In an effort to overcome the barriers to data sharing and to facilitate collaboration between law enforcement agencies involved in this project, the TPD has developed a generic Intergovernmental Agreement (IGA) that is signed between different law enforcement agencies participating in the project. This IGA was condensed from MOU's (memorandum of understanding), policies and agreements that previously existed in various forms between numerous agencies. In order to overcome the barriers caused by concerns over data privacy and security, we have taken certain necessary measures. The data shared between agencies contains only law enforcement data and is available only to individuals screened by these agencies using TPD Background

Check, Employee Non-Disclosure Agreement (NDA) and the TOC (terminal operator certificate) test. In addition, all law enforcement data in the University of Arizona's AI Lab reside behind a software firewall and in a secure room accessible only by activated cards to those who have met the above criteria.

2.2 Identity Resolution

Identity resolution is critical to various governmental practices ranging from providing services to citizens to enforcing homeland security. The task of searching for a specific identity is difficult because multiple identity representations may exist due to issues related to unintentional errors and intentional deception. To the best of our knowledge, there are few solutions proposed for this problem. We have proposed a probabilistic Naïve Bayes model that improves existing identity matching techniques in terms of effectiveness.

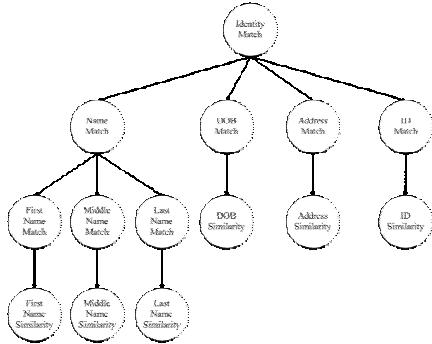


Figure 1. A probabilistic model for identity matching.

The proposed model includes four identity features that were found more reliable than others for identity matching in a previous case study. The model is mainly innovative in two aspects. First, this model associates a probability rating to every possibly matching identity. Users can understand probability ratings easily by converting them into human belief. Second, our proposed model reduces human intervention in learning by using a semi-supervised technique. Experiments show that our proposed model performs significantly better than the exact-match based matching technique. In addition, our model greatly reduces the efforts of manually labeling training instances by employing a semi-supervised learning approach. This training method outperforms both fully supervised and unsupervised learning. With a training dataset that only contains 20% labeled instances, our model achieves a performance comparable to that of a fully supervised learning [2].

2.3 Mutual Information to Identify Target Vehicles

In recent years border security has been identified as a critical part of homeland security. The Department of Homeland Security (DHS) monitors vehicles entering and leaving the country, recording their license plates with a date and time of entry using license plate readers. These thorough checks are done for vehicles on watch lists (target vehicles) and on random vehicles as well. This process is time consuming and if the waiting times become too long, the flow of people, vehicles, and commerce is impaired. So, CBP agents are under pressure to balance security needs with

efficiency. One of the aims of this study is to help CBP agents identify better quality target vehicles.

CBP agents believe that vehicles involved in illegal activity (especially smuggling) operate in groups. If the criminal links of one vehicle in a group are known, then the group's crossing patterns and frequency can be used to identify other partner vehicles. We perform this association analysis by using mutual information to identify pairs of vehicles crossing together and potentially involved in criminal activity. Our previous study [3] had found that the use of MI was a promising solution to this problem. In this study we modify the MI measure to incorporate domain heuristics. Domain experts (CBP agents, police detectives and analysts) suggest that groups of criminal vehicles may cross at certain times during the day to try and evade inspection. We use law enforcement information from border-area jurisdictions to identify times that criminal vehicles prefer and attempt to incorporate this knowledge in the MI formulation using conditional probability.

We find that mutual information can be used to identify high quality potential target vehicles at the border. In addition, an initial experiment showed that the mutual information measure modified to include domain heuristics such as time of crossing performs better than classical mutual information in the identification of criminal vehicles. The method can be used to assist CBP agents to perform their functions both effectively and efficiently. In the future, we plan to incorporate other domain heuristics in the mutual information formulation. We also plan to have domain experts from Customs and Border Protection validate our results and operationalize them.

3. CONCLUSIONS

Through the accomplishments of the COPLINK center, we can provide law enforcement and intelligence communities with the most recent research technology advancements in a user-friendly manner that can be easily used by their staff to improve their job performance. Our research accomplishments also help the advancement of research in areas such as criminal network analysis and visualization, identity resolution and applications of mutual information.

4. ACKNOWLEDGMENTS

We would like to thank the Tucson Police Department, Tucson Customs and Border Protection and members of the University of Arizona's Artificial Intelligence Lab.

5. REFERENCES

- [1] Chen H., D. Zeng, H. Atabakhsh, W. Wyzga, and J. Schroeder, "COPLINK managing law enforcement data and knowledge," Communications of the ACM, vol. 46, pp. 28-34, 2003.
- [2] G. Wang, "A Multi-Layer Graphical Model for Approximate Identity Matching," In Proceedings of the AIS Americas Conference on Information Systems (AMCIS 2005), Omaha, NE, USA, August 11-14, 2005.
- [3] Kaza, S., Wang, T., Gowda, H and Chen, H. Target Vehicle Identification using Mutual Information. In Proceedings of the 8th International IEEE Conference on Intelligent Transportation Systems, Vienna, Austria, 2005.

SESSION 2B

CITIZEN PARTICIPATION 1

Moderator

Teresa Harrison, University at Albany/SUNY, USA

Titles and Authors

Should E-Government Design for Citizen Participation? Stealth Democracy and Deliberation
Muhlberger, Peter

Converting Online Public Legal Information into Knowledge: “ABC del Diritto” an Italian e-Government Citizen-oriented Service
Biasiotti, Maria Angela; Nannucci, Roberta

Web Portal Implementation to Support Public Participation in Transportation Decision Making
Nyerges, Tim; Brooks, Terry; Jankowski, Piotr; Rutherford, G. Scott; Young, Rhonda

Should E-Government Design for Citizen Participation? Stealth Democracy and Deliberation

Peter Muhlberger

Univ. of Pittsburgh; University Center for
Social and Urban Research
121 University Place
Pittsburgh, PA 15260
001-412-648-7099

pmuhl830@gmail.com

ABSTRACT

Cyberoptimists have heralded an age of citizen engagement enabled by electronic technologies that allow widespread citizen input in government decision making. In contrast, influential political scientists maintain that the preponderance of citizens quite reasonably wish to avoid political participation and that involving citizens could have very negative consequences for governance. In their widely-read book, *Stealth Democracy*, Hibbing and Theiss-Morse seek to show that much of the American public desires "stealth democracy"—a democracy run like a business by experts with little deliberation or public input. The authors maintain that stealth democracy beliefs are due to reasonable apathy rationales and that a more engaged democracy is simply of no interest to the public. This paper introduces an opposing "parochial citizens thesis" that suggests that stealth democracy beliefs may be driven by socially problematic beliefs and orientations, including reverence for authority and an incapacity to take other political perspectives. These views are rooted in simplistic conceptions of human agency and political leadership that might be ameliorated through deliberation. This paper examines survey and experimental data from the National Science Foundation / Information Technology Research funded Virtual Agora Project. The data comprise a representative sample of 568 Pittsburgh residents, who participated in face-to-face and online deliberations. Using OLS regression with cluster-robust standard errors, the paper finds that stealth democracy beliefs are explained by beliefs and orientations consistent with the parochial citizens thesis. It also finds that *online* democratic deliberation significantly ameliorates key stealth democracy beliefs and some of the factors that lead to these beliefs. Contrary to the stealth democracy thesis, e-government efforts to stimulate citizen deliberation may have positive consequences.

Categories and Subject Descriptors

K.4.0 [Computers and Society: General]

General Terms

Management, Performance, Experimentation, Human Factors.

Keywords

Political Apathy, Stealth Democracy, Political Participation, Political Discussion, Democratic Deliberation, Online Deliberation, Human Agency.

1. INTRODUCTION

Cyberoptimists hold that information technology (IT) will appreciably reduce political ignorance and apathy and enable citizens to provide substantial input into government decision making [1, 4, 11, 23, 27]. Researchers and enthusiasts express the hope of benefits from using electronic technologies for more deliberative input into government decision processes [17, 30]—input involving discussion between citizens. While IT enthusiasts and researchers embrace the prospect of greater citizen engagement and deliberation through technology, many political scientists have cobbled together a conception of the public that recommends against civic engagement efforts. In this view, the public has a strong and reasonable desire to not trouble itself with political matters and efforts to involve the public, particularly in deliberation, could quite adversely affect governance, perhaps delegitimizing the political system.

Hibbing and Theiss-Morse [9] find that 93.5% of a representative survey sample of the American public agree with one or more of three statements describing what they call "stealth democracy" beliefs. These are statements that express intense impatience with debate and compromise among political leaders and a desire to have government run by successful business leaders or unelected independent experts.

In addition, Hibbing and Theiss-Morse shape their various findings into a book-length argument against prescriptions to engage the public more deeply in politics, particularly prescriptions for deliberative involvement. Their "stealth democracy" thesis holds that much of the public is uninterested in politics, dislikes conflict, and believes that there is wide consensus on political goals. Because the public believes there is wide consensus, it does not see the point of disagreement and conflict in politics. The authors maintain that more deeply involving such a public in political life is a prescription for frustration, distrust, and delegitimization of the political system.

The stealth democracy thesis has been well received by many political scientists. The book received favorable reviews by such luminaries as Robert Shapiro [28] and has become a mainstay of many college courses in political science and public opinion—

Google finds 324 web documents that mention the book in relation to the word "course." Google Scholar also finds 166 references to the book and related academic papers, and references to "stealth democracy" occur in 37 papers presented at the 2005 Annual Meeting of the American Political Science Association. The concept has also come into use outside academia, as indicated by a Google search finding 503 references to the term "stealth democracy" exclusive of references to courses. Some lines of inquiry within mainstream political science are more favorable to deliberation—such as Putnam's social capital approach [21] or Fishkin and Luskin's Deliberative Polling work [12]. Nevertheless, the stealth democracy thesis has made significant inroads in the mainstream of the profession in the few years since the book was published.

An important aspect of Hibbing and Theiss-Morse's position is normative. Their overarching concern is with insuring the stability and legitimacy of the political system. Consequently, in their chapter of prescriptions, they do not recommend ways to reverse political disinterest or conflict aversion, which they do not see as injurious to system legitimacy. The book depicts political ignorance and disinterest as "perfectly understandable" (p. 134) and discomfort with conflict in political discussion as "avoiding a distasteful activity," a dislike that makes "perfect sense" (p. 10). People are described as naturally more interested in their everyday lives than in politics.

Thus, Hibbing and Theiss-Morse do not consider stealth democracy anti-democratic but simply realistic in light of the public's reasonable preference to be politically uninvolved. Only one matter disturbs the authors—the public's false belief in a political consensus—because they fear false perceptions of consensus may delegitimize the political system. False consensus beliefs create unrealistic expectations that leaders can readily act with little debate or compromise. The authors recommend such beliefs be addressed with an intensive educational effort.

To the stealth democracy thesis, this paper opposes the "parochial citizens thesis." This thesis claims that many people have simplistic understandings of human agency. These understandings result in an inability to conceptualize complex systems of governance and an inability to take alternative political perspectives. Such underdeveloped reasoning about politics leads people to falsely believe in political consensus and to embrace undemocratic forms of governance, specifically authority-driven stealth democracy.

Contrary to the stealth thesis, the parochial citizens thesis suggests deliberation could be beneficial. Educational experiences might ameliorate such "parochial reasoning" by calling on people to refine their thinking about politics. Deliberation in particular could both help clarify that reasonable people hold a diversity of views and exposes discussants to complex processes of decision making that might undermine stealth democracy beliefs. With such deliberative methods as the National Issues Forums and Deliberative Polling, it is commonplace for practitioners and researchers to find that participants engage in respectful and thoughtful discussions of the issues as well as their differences [8, 19]. Online deliberation in particular may be beneficial because the deindividuating effects of online environments could encourage people to think more as individuals, generating more disagreement [18].

While this paper cannot address every aspect of the parochial citizen thesis, it will test: a) whether stealth democracy beliefs are

grounded in unreasonable and socially-problematic views and orientations as predicted by the thesis and b) whether online deliberation helps to ameliorate stealth democracy beliefs and some of the problematic views and orientations that contribute to these beliefs. This paper examines these hypotheses with data from a National Science Foundation-funded study of democratic deliberations involving 568 Pittsburgh residents selected by random digit dialing. The findings are consistent with the parochial citizens thesis, suggesting that e-government efforts to encourage citizen participation, particularly deliberative participation, will not run contrary to a reasonable public desire to be politically uninvolved and may have positive benefits in cultivating a more civically-minded public.

2.PAROCHIAL CITIZENS—A THEORY

The idea of parochial citizens was inspired by the implications of linear reasoning, a particular type of causal reasoning, for political understandings. Linear reasoning is a concept from Rosenberg's [24, 25] cognitive developmental theory and research. The reader need not fully subscribe to this cognitive developmental theory, but only recognize that linear reasoning provides a coherent description of a type of reasoning that people might exhibit on certain topics, particularly political topics about which they have limited understandings.

Rosenberg's [24, 25] cognitive developmental theory and research suggests that many adults understand their world through "linear reasoning." In linear reasoning, people understand causality by focusing on an anchoring entity from which effects flow in a simple, direct manner. Linear reasoners conceptualize causal systems as simple linear chains involving single causes for any given effect. Unlike Rosenberg's systematic reasoners, linear reasoners do not adequately understand systems, which have multiple causes to an effect, feedback loops, and systemic properties such as system goals and principles of operation.

Whether or not the explanation of the public's reasoning about politics has a cognitive developmental component, most Americans' attention to and understanding of political matters are so limited [5, 7, 10, 16] that it would be surprising to find systemic understandings of politics. Linear reasoning might appear when people's knowledge of a topic is insufficient to rise to a systemic level.

The logic of their reasoning has implications for how linear reasoners understand human and political agency [14], and these implications give rise to the parochial citizen worldview. A linear thinker can only conceptualize government as under the control of a single strong leader. The parochial worldview must further accommodate itself, in the West, to the knowledge that the political system is democratic. I propose it does so by stipulating a monolithic public opinion that is interpreted by a strong leader with special knowledge of the public, such as the President, who in turn directs the government to carry out the wishes of "The Public."

The parochial worldview also involves ethical judgments that evoke emotion and motivation. An organization under the full control of a monolithic will is a direct indicator of the moral qualities of its leader. Given that an undifferentiated will directly manifests itself in the actions of government, good actions must indicate that the will is all good and bad actions must indicate it is all bad. The logic of the parochial worldview leads to a morally totalizing comprehension of government—government is either

all good or all bad. Parochial citizens, then, view government in black or white terms, usually forming an entirely positive normative stance toward the government.

The parochial cognitive model of government poorly reflects reality and must therefore be maintained in the face of contradictory information. Parochial citizens will be motivated to defend their cognitive model because of its all-positive normative content and their inability to see any conceptual alternative. For example, a challenge to the belief in the monolithic quality of the public will is also a challenge to the possibility of democracy, because no other kind of democracy can be conceived. To the extent that they become aware of conflicting views in the public, and surely they must be aware of some conflict, they may dismiss it as representing "un-American" (or "un-British", "un-French", etc.) viewpoints—that is, by redefining the "true" public to not include the dissenting views. Similarly, parochial citizens will be motivated to reject negative information on a government they view favorably.

3. STEALTH DEMOCRACY AND THE PAROCHIAL CITIZEN

The parochial citizen should be predisposed toward stealth democracy beliefs. To the extent that they view the political system as having any good effects, those with the parochial worldview are inclined to believe that all aspects of the political system are good. Dissent, then, goes against the single, all-good will that constitutes the political system. Elites are seen as essential interpreters of the "true public will." Thus, parochial citizens should be inclined to prefer a political system without debate or compromise run by elites who interpret and implement a common public will—hallmarks of stealth democracy beliefs.

Between the abstract logic of linear reasoning on the one hand and stealth democracy beliefs on the other are a range of intermediate attitudes that should be characteristic of parochial citizens—false beliefs in a public consensus, fear of conflict, reverence of authority, incapacity for social perspective taking, and passivity with respect to cognition. Linear reasoning inclines people toward these attitudes and these attitudes in turn stoke stealth democracy beliefs. Parochial citizens' belief in a monolithic public will naturally lead to a false belief in public consensus on policy. As already noted, however, parochial citizens may be somewhat conflicted between their desire to believe in a mythic consensus and awareness of dissent in the real public. Parochial citizens may be especially troubled by dissent precisely because it conflicts with their notion of democracy. Paradoxically, the parochial citizen may therefore be driven to embrace stealth democracy both out of a belief in an abstract public consensus *and* out of fear of concrete conflict (this will be presented as the variable Expect Conflict in data analyses later in this paper).

The parochial citizen also embraces hierarchy in government, a hierarchy dominated by strong leaders. Parochial citizens do not understand systems of checks and balances, which are guided by system principles and goals. Moreover, parochial citizens feel a strong normative call to defend or revile groups and organizations they understand in black and white terms. Thus, parochial citizens are drawn to positive views of social hierarchy and authority, perhaps including vertical collectivism (the variable VC, belief that individuals should suppress their wishes and goals on behalf of their group-oriented roles), right-wing authoritarianism (RWA, obedience to authority and punitive

attitudes toward the disobedient—I have removed the traditionalism component), and social-dominance orientation (SDO, belief that some social groups are better than others and should dominate). An extensive literature links these authority attitudes to socially problematic outcomes such as prejudice, irrationally punitive political attitudes, and close-mindedness [2, 13, 31].

Because they are apt to value a monolithic public will, parochial citizens should be disinclined toward *political empathy*—taking the political perspective of other racial and class groups and of those who disagree with themselves politically. Likewise, they should be inclined toward *naive realism*—an incapacity to understand political disagreement because of an inability to take the perspective of the dissenter. Naive realists see their own perspective as self-evident and those of dissenters as incomprehensible. Consequently, they rationalize disagreement as due to lack of effort by dissenters or due to their irrationality or ill-intent. Those low in political empathy and high in naive realism should be particularly susceptible to belief in a false consensus and may therefore be more amenable to stealth democracy.

The parochial citizen may also possess certain cognitive dispositions. The parochial worldview involves a serious oversimplification of reality, which means consistency is only possible by ignoring many facts, and it reinforces an unquestioning attitude by reviling dissent itself. Thus, parochial citizens should be inclined toward moderately low need for cognition (NFC, a self-report measure of enjoyment of thinking) and toward high need for structure-order (NFS, a desire for certainty and order). Those low in NFC and high in NFS might prefer a stealth democracy because these dispositions play into authority attitudes, false consensus beliefs, and political perspective taking. They might also directly prefer such a democracy because they expect the public, like themselves, to prefer not to exercise their cognition or address uncertainty.

4. METHOD

4.1 Participants

Knowledge Networks (KN), an outside firm noted for its sampling work on academic deliberation projects, conducted the recruitment for this study. Of a sample of 6,935 Pittsburgh city residents (defined by zip code area) who could be reached via random digit dialing (RDD), 22% agreed to participate in this research and took a phone survey. Sampling differed from KN's typical methodology on other deliberation projects in that it did not utilize quota sampling to make demographic statistics more representative of the population as a whole. Thus, the sample accurately reflects who would come to this deliberation without demographic oversampling. The sample better generalizes to what it would be if deliberation were a more widely used process of government, because cost and legal requirements would likely prevent quota sampling. Also, it avoids the concern that those who come to a deliberation after extensive oversampling may be atypical of their demographic.

Of recruits who agreed to participate, 37% or 568 people showed for the Phase 1 on-campus deliberation. Knowledge Networks succeeded in phone-interviewing 463 of the 568 study participants before they came to their on-campus day of deliberation. A modest response rate was expected because recruits were asked to participate in a series of online deliberations that would take most participants eight-months to complete and which they could join

only by coming to the initial on-campus, all-day deliberation. The final participation percentages are not, however, incomparable to that of another substantial long-term deliberation study, Vincent Price's Electronic Dialogue Project at the Annenberg School of Communication [19, 20]. This project started with an effective sample of the population from which its discussants were drawn of about 3,686 [20]. The number of people who ever participated in any discussion over the course of the year is 543, and the average number of people who participated in a given discussion was 305 [19]. Ultimately, the response rates are modest. Comfort can be drawn from several considerations: a fair similarity to population demographics, the fact that the sample represents people who might be expected to participate in longer-term deliberations, and the objective of this research which is experimental and focused on psychological processes that should be universal.

Despite a strict RDD sample and modest response rate, the participants in this project reasonably matched the Pittsburgh city population on most demographic criteria. The sample was 77% Caucasian and 18% African-American, compared with CPS population benchmarks for the relevant zip codes of 75% and 20%, respectively. Fifty-six percent of the sample was female, compared with 53% for the population. Twelve percent of the sample was 18-29 years old, 22% 30-44 years old, 26% 45-59, and 27% 60+. This compares with population values of 26%, 20%, 26%, and 27%. The elderly and thirty-somethings are accurately represented, the young are underrepresented, while mid-life adults are overrepresented. Average age, however, is the same as for the population. Perhaps the greatest departure from population values is for education, which, as expected, is greater than for the population. Median education is "Some College" for both the sample and the population. Lower educational categories, however, are underrepresented, with 10% of the sample having less than a high school education and 14% having just a high school education, compared with 16% and 31% for the population. Nevertheless, the sample does contain the full range of educational levels.

Phase 2 of the project, the eight-month at-home online deliberations, was intended to include 410 of the original 568 participants who were selected to receive a computer. Substantial participant drop-off occurred by Phase 2 of the project, with response rates to questionnaires in the early part of Phase 2 dropping to about 230. Drop-out was perhaps driven in part by participant frustration with software and hardware problems and disappointment with the quality of the computer equipment provided as an incentive. The project's capacity to purchase high quality equipment and to address other problems was constrained by the resources allotted for social research on the project.

Pittsburgh is an ethnically and class diverse community with a city population of 334,583 and over one million including surrounding areas, according to the 2000 Census. Neighborhoods range from suburb-like residential areas to areas of urban poverty. Although Pittsburgh is known to have a moderately high quality of life for a city its size, people intimately involved with public life in the city do not believe this leads to either an especially high level of political involvement or non-contentious public dialogue.

4.2 Materials and Procedures

Knowledge Networks obtained phone numbers for households in the City of Pittsburgh from a random digit dial (RDD) sample.

Where numbers appeared in a reverse directory, the household was sent an advance letter on Carnegie Mellon University stationery describing the study and indicating that the household would be contacted shortly. A Knowledge Networks phone center called households in the RDD sample and requested the household member with the most recent birth date. Both the letter and the call center indicated that in exchange for participation in the study, participants would have a four out of five chance of receiving a Windows computer and eight months of ISP service. The remainder would receive \$100. Those who received a computer would be expected to participate in a longer-term online deliberation from home that would require six hours of discussion over eight months. People who agreed to participate were given a short phone-based survey of their demographics and a few policy attitudes, and they were scheduled for a one-day, eight hour on-campus deliberation. Participants were asked to come to a randomly-chosen day from the deliberation schedule, which spanned three weeks in July, including many weekends and weekdays.

Deliberations were held with up to 60 participants daily. After informed consent and a brief training session, participants took a web-based pre-survey. Next, they were given a 40 minute "library session" to learn more about the four policy topics, a break, 90 minutes for "deliberation" (face-to-face, online, or individual contemplation, depending on condition), and lunch. The library session, break, and deliberation (same condition as before) were repeated in the afternoon, and this was followed by the second survey. In addition to the experiment with type of deliberation, another experimental condition involved either receiving or not receiving reminders of citizenship. In the citizenship condition, participants were reminded to think like citizens in a brief "talking-head" ahead of their deliberations (the non-citizen condition involved a different talking-head), their rooms had an American flag, and they were given name tags with American flags and the word "Citizen" preceding their names.

4.3 Measures

4.3.1 Apathy Rationales

The apathy rationales were each measured with multiple questions. Apathy rationale questions appeared in random order. All question responses were measured on 7-point Likert scales. A sample question is: Conflict Averse (Phase 1 post-deliberation survey)—"When people argue about politics, I feel uneasy and uncomfortable." Note that conflict aversion involves a slight rewrite of the Hibbing and Theiss-Morse question so it would fit better into a set of Likert questions. It was joined by a companion reversed question.

One apathy rationale occurred in the pre-deliberation questionnaire, False Consensus—"Thinking about the American people, what portion of Americans do you believe think 'MostImpProblem' is the single biggest problem facing the country today?" and "What portion of Americans do you believe basically agree with you on what should be done about 'MostImpProblem?'". The survey system replaced MostImpProblem with the most important problem facing America that the participant had earlier identified. The 11-point response scale had labels: No Americans, Half of All Americans, All Americans. This response scale has an objective interpretation, unlike Hibbing and Theiss-Morse's "very few, some, most" Americans scale. I also added another pre-deliberation measure of expected unproductive conflict: Expect

Conflict—"Overall, what portion of discussion in your discussion group do you anticipate will involve unproductive conflict?" (11-pt. scale anchors: None of the Discussion / Half of the Discussion / All of the Discussion).

4.3.2 Authority Attitudes and Cognitive dispositions

Most of these were measured using short versions (4-6 items) of scales widely used and accepted by political and personality psychologists and can readily be found in a search of PsychInfo. This includes social dominance orientation (SDO)[31], right-wing authoritarianism (RWA)[2], vertical collectivism (VC)[33], need for cognition (NFC)[3], and need for structure-order (NFS)[15]. One novel measure is naive realism, the idea for which was suggested by Ross [26]. It involves such questions as: "I can understand why people who disagree with me politically believe what they believe." and "People who disagree with me politically seem to have an agenda." The second novel measure is political empathy. The measure involved rewriting the Interpersonal Reactivity Index (IRI) questions pertaining to empathic perspective taking [6] so that they focused on politically-relevant rather than interpersonal perspective taking. These include questions such as: "If I'm sure I'm right about a political issue, I don't waste much time listening to other people's arguments." and "I sometimes find it difficult to see political issues from the point of view of people in other social classes."

5.RESULTS

5.1 The Contentious Nature of the Issues

The topic of deliberation was Pittsburgh public school consolidation and three related policies. Because of population decline, Pittsburgh public schools had a substantial and expensive excess of seating capacity in schools. The issue is contentious, pitting parents against taxpayers and neighborhoods against the School Board. Fifty-four percent of participants reported that the issues directly affected them or their families.

5.2 Explaining Stealth Democracy

Table 1 shows regressions of stealth democracy on three models. All analyses are conducted with robust errors that account for discussion group error covariance, because deliberation in groups may have affected some of the variables involved. The model in the first column after the variable names (henceforth Column 2) seeks to reproduce Hibbing and Theiss-Morse's regression, except that rather than creating a single indicator called "negative view of disagreement" that averages false perceptions of a public consensus, aversion to conflict, and political interest, these three variables are each entered separately. Averaging these questions runs contrary to the authors' theoretical discussion and obscures important differences in the effects of the variables. Column 2 shows that political (dis)interest plays no significant role in explaining stealth democracy beliefs, while false consensus perceptions have 2.4 times the effect of aversion to conflict. Note that continuous variables were put on seven-point scales to insure comparability of coefficients. With addition of yet other control variables in Column 3, conflict aversion proves non-significant, suggesting that it may have merely a spurious or indirect relationship with stealth democracy beliefs. Despite their central role in the stealth democracy thesis, personal discomfort with conflict and political disinterest are not the dominant factors in explaining stealth democracy beliefs.

Table 1. OLS Regressions of Stealth Democ. on Three Models

Independent Variables	All non-dichot. vars on 7-pt scales. Unstandardized Coef. (Cluster-Robust s.e.)		
	Parochial Cit.		.43***(.06)
VC		.21*** (.05)	
RWA		.17** (.06)	
SDO		.06 (.05)	
False Consen.	.22***(.04)	.17***(.04)	
Exp. Conflict		.13*** (.04)	
Naïve Realism		.17** (.06)	
Social Empty		-.05 (.07)	
NFC		.02 (.07)	
NFS		-.06 (.06)	
Conflct Avers.	.09** (.03)	.05 (.03)	.04 (.03)
Political Inter.	-.04 (.04)	-.03 (.04)	.03 (.04)
Liberal	-.12*** (.04)	-.02 (.04)	-.07* (.04)
Democrat	-.16 (.11)	-.16 (.10)	-.16 (.10)
Republican	-.11 (.18)	-.24 (.17)	-.31* (.16)
Education	-.27*** (.04)	-.20*** (.04)	-.20***(.04)
All analyses also control for income, ethnicity, gender, age, & constant (not shwn)			
R ² ; s.e.	.25; 1.09	.34; 1.03	.29; 1.06

Note: N=558 throughout (loss of 10 observations due to non-response). All F-values<.0001. *** is p<.001; ** is p<.01; * is p<.05; † is p<.10 All p-values are robust and account for non-independence of errors by discussion group. P-values reported are one-sided for all non-demographic variables with coefficients in the expected direction.

Column 3 of Table 1 displays the full model derived from the parochial citizen thesis, along with the Hibbing and Theiss-Morse model. The Column 3 model is superior to the Column 2 model in terms of R² and standard error. The only variable from the Hibbing and Theiss-Morse theoretical model that remains significant in Column 3 is false consensus beliefs. The most potent variable in the table is vertical collectivism (VC) and the combination of VC and RWA, both authoritarian beliefs, dominates the effects. As predicted by the parochial citizen thesis, both false beliefs in an abstract public consensus and expectations of unproductive conflict in the concrete deliberations contribute to stealth democracy beliefs. This poses a paradox for the stealth democracy thesis.

Column 4 of Table 1 tests the possibility that a single composite indicator combining all the views and orientations of the parochial citizen mentality might do well in explaining stealth democracy beliefs. The composite is simply a weighted average of the variables, with weights determined by an exploratory factor analysis fitting these variables to one factor. The composite quite potently explains stealth democracy beliefs. While the amount of explained variance is lower than for Column 3, this may be related to simply having fewer variables with which to overfit the dependent variable.

5.3 Effects of Online and F2F Deliberation

Table 2 presents results indicating that deliberation helps ameliorate stealth democracy beliefs and some of the variables feeding into stealth beliefs. Only two of the nine variables underlying the parochial citizen thesis were available for consideration. A decision was made to not include pre- and post-deliberation measures in Phase 1 for most indicators because of concern that pre-measures administered the same day as post-measures would prove reactive. Instead, it was anticipated that

post-measures would be collected during Phase 2. Regrettably, Phase 2 experienced considerable respondent drop-out, for reasons previously discussed. In addition, Phase 2 began later than expected because of software issues, creating a remove of several months between measures collected in both Phases. Therefore, change between Phases 1 and 2 will, with the exception of the crucial stealth democracy beliefs variable, not be considered because of small sample size (hence low statistical power), possible weakening of effects over time, and the possibility that intervening events may have influenced the variables.

One of the nine parochial citizen variables was collected post-discussion in Phase 1: vertical collectivism (VC). It is therefore possible to determine whether VC was significantly larger for those who deliberated than those who did not. Column 2 of Table 2 (the first column of results) shows an ANOVA-equivalent regression. The constant indicates the constant for VC in the excluded condition: Control X No Citizen—that is, no discussion and no reminders of citizenship. Coefficients for the other conditions indicate deviation from this overall constant. Thus, for example, the mean level for Online X Citizen is .73-.33 or .40. Column 2 shows that the two online discussion conditions had very significantly lower levels of post-discussion VC than the Control X No Citizen condition (they are also significant if the contrast condition is *both* control conditions— $p=.03, .01$). Lower levels of VC were expected to be most noticeable for the Online X No Citizen condition, but it appears that both online conditions contributed equally to reduce VC. The findings on VC are quite definitive—deliberation reduces vertical collectivism, which is one of the primary contributors to stealth democracy beliefs.

Table 2. OLS Regressions Showing Effects of Deliberation on Outcome Variables

	Dependent Variables		
	VC (post-delib.)	Change in Exp. Confl.	Change in Stealth Bel's ^a
Independent Variables	All non-dichot. vars on 7-pt scales. Unstandardized Coef. (Cluster-Robust s.e.)		
Online X Citiz.	-.33** (.13)	-.74† (.54)	-.45† (.34)
OnlineXNo Cit	-.32** (.11)	-1.23** (.46)	-.60* (.34)
F2F X Citiz.	.07 (.15)	-.78† (.52)	-.69* (.30)
F2F X No Cit.	-.01 (.15)	-1.14* (.62)	-.54* (.32)
Ctrl X Citiz.	-.20 (.14)	-.41 (.52)	-.51 (.44)
Ctrl X No Cit.	See Cons.	-.32 (.47)	-.35 (.26)
Education	-.14***(.03)	.12 (.09)	.09 (.07)
Income	.05 (.03)	-.15† (.08)	-.03 (.07)
Age	.10* (.04)	-.28** (.10)	.05 (.07)
African-Amer.	.10 (.11)	.82** (.29)	-.18 (.25)
Male	.22* (.09)	.23 (.18)	-.18 (.25)
Constant	.73***(.17)	N / A	N / A
N; R ² ; s.e.	556; .08; .98	559; .22; 2.4	229; .05; 1.3

Note: All F-values<.0001. *** is $p<.001$; ** is $p<.01$; * is $p<.05$; † is $p<.10$. All p-values are robust and account for non-independence of errors by discussion group. P-values reported are one-sided for all non-demographic variables with coefficients in the expected direction.

^aA subset of two stealth democracy variables (see text)

Expectations of unproductive conflict can be compared with a post-deliberation Phase 1 measure of perceptions of conflict. Perceptions of conflict was collected from deliberators by asking about how much conflict they perceived in their discussions. This

variable was collected from the control group by asking participants at the end of Phase 1 how much conflict they would anticipate in a discussion. (This "perceived conflict" variable differs, however, from the pre-deliberation expected unproductive conflict variable in that it does not use the word "unproductive." Statistical evidence from the survey indicates that the two variables are closely related. Indeed the post-deliberation survey included a perceived *unproductive* conflict question for discussion group members that is highly correlated with the post-deliberation perceived conflict question [$p=.59$.]) Column 3 of Table 2 shows a regression of the *change* in perceived conflict (post-deliberation perceived conflict minus pre-deliberation expected unproductive conflict) on the experimental conditions. Coefficients of the experimental conditions indicate the amount by which post-deliberation perceived conflict changes from pre-deliberation expected conflict in that condition. Changes are quite substantial and negative, indicating large declines in perceived conflict, with significant effects in two experimental conditions. Another regression (not depicted) asked whether this change was significantly more negative in the discussion than control conditions, it was ($\beta=-.63$, $p=.01$, for a variable coded 1 for discussants and 0 otherwise). (An examination of the post-deliberation perceived *unproductive* conflict variable shows highly significant decreases in this variable relative to the pre-deliberation variable for all discussion conditions ($p<.012$ for all). This analysis has the weakness that it does not include observations in the control conditions.)

Column 4 of Table 2 shows significant reductions in stealth democracy beliefs in three of four discussion conditions with a trend in the fourth. The stealth democracy variable here is an average of only two of the four stealth variables. Indications in correlation patterns and means suggests that there may be some difference between two stealth democracy variables that ask whether debate and compromise should be cut short in government and two other variables that ask whether government should be run by experts and business leaders. The citizen-to-citizen deliberation in the current study can be expected to have greatest effects on perceptions of debate and compromise, not the value of business leaders and experts in government. Also, 73% of participants in the current study disagreed with one or both of the questions about the desirability of a government run by business leaders and experts. If deliberation reduces stealth democracy beliefs, there would be little room to register reductions on these variables. In contrast, 82% of participants agreed with one or both questions indicating the desirability of reducing debate and compromise, allowing considerable room for improvement. Not surprisingly, deliberation has no significant effect on the business leaders and experts questions, but it does have effects on the debate and compromise questions, upon which I focus here. The change reported in Table 2 is measured as the difference between Phase 2 two-variable stealth democracy beliefs and Phase 1 pre-deliberation two-variable stealth democracy beliefs. The Phase 2 beliefs were measured shortly after the start of Phase 2. The strongest effect was in the f2f X citizen condition, followed by the online X no-citizen condition, which was expected to be the strongest effect.

6.DISCUSSION OF FINDINGS

Researchers and practitioners in e-government are optimistic about the benefits of communication technologies for democracy. In contrast, many political scientists entertain the stealth

democracy thesis that most of the public desires a democracy with little debate, compromise, or public input run by experts and business people. Indeed, Hibbing and Theiss-Morse find considerable agreement in the public with questions tapping stealth democracy beliefs. Their findings lead these authors to believe that encouraging public participation would either be irrelevant because of reasonable public disinterest or potentially trigger adverse consequences such as system delegitimization.

This paper proposes a different interpretation of the finding that Americans embrace stealth democracy beliefs. It stipulates that these beliefs are rooted in a "parochial citizen worldview" involving a set of socially problematic views and orientations and that this syndrome can be ameliorated by involving people in online political deliberation. The views and orientations include false consensus beliefs, fear of conflict, strong pro-authority attitudes, incapacity for social perspective taking, and dispositions to cognitive lethargy.

The paper's findings are consistent with the parochial citizen thesis. The nine parochial citizen views and orientations prove to be a far better explanatory model than Hibbing and Theiss-Morse's original model that focuses on false consensus beliefs, political disinterest, and aversion to conflict. Indeed, the latter two variables prove non-significant, challenging Hibbing and Theiss-Morse's interpretation of stealth democracy beliefs as rooted in understandable political disinterest and aversion to conflict. Strong pro-authority beliefs, associated in the literature with prejudice and irrationally punitive attitudes, are the most potent explanation of stealth democracy beliefs. Also, the Hibbing and Theiss-Morse interpretation cannot explain why participants in the present study embraced stealth democracy both out of a false belief in a consensus and fear of conflict. The parochial citizen thesis explains how the same people can both believe in an abstract consensus and fear actual conflict.

Finally, the paper reveals that democratic deliberation mitigates two of the key components of stealth democracy beliefs and some of the views and orientations behind these beliefs. Deliberation reduces post-deliberation attitudes, including stealth democracy beliefs as well as vertical collectivism and perceptions of conflict—potent explanations of stealth democracy beliefs. While not examined here, the data on which the current paper is based clearly show that deliberation does not decrease confidence in government, alleviating concern about system delegitimization.

7. IMPLICATIONS FOR DIGITAL GOVERNMENT

The implications here for digital government are positive. This paper introduces a theoretical and empirical response to the stealth democracy thesis—a claim widely promulgated in political science that citizens neither want nor would benefit from greater engagement. The paper suggests a counter-interpretation of stealth democracy findings as rooted in parochial citizens whose overly simple understandings of government might threaten democracy. In this interpretation, citizen engagement and discussion may be crucial to ameliorating these simplistic understandings and the socially harmful beliefs and orientations to which they give rise. Results from the present study show that, in particular, two key factors play no significant direct role in explaining stealth democracy—personal aversion to conflict and political disinterest. The stipulated relationship between these two factors and stealth democracy beliefs account for Hibbing and

Theiss-Morse's conclusion that citizens do not want more engagement. Showing that a direct relationship does not exist helps undermine this conclusion. Instead, stealth democracy appears rooted in a syndrome of authoritarianism, poor socio-political perspective taking, and cognitive lethargy. This syndrome can be understood as arising from inadequately developed understandings of political agency—of leadership and the dynamics of organizations. Deliberation might directly undermine core beliefs behind this parochial citizen mentality—by demonstrating to participants that reasonable and patriotic people can disagree on the issues and that the public can amicably and intelligently arrive at a solution without political leaders imposing a "consensus." Deliberation should and does undermine stealth democracy beliefs and some of the factors that lead to these beliefs. The theory and findings here strongly indicate that, yes, government should design for citizen participation, especially deliberative participation.

Moreover, the findings in particular indicate that e-government deliberative initiatives would be worthwhile. A common perception about deliberation practitioners (at the Deliberative Democracy Consortium, personal communications) is that face-to-face (f2f) deliberation is vastly superior to online deliberation. The Kettering Foundation, for example, has long refused to entertain online deliberation because of a conviction that such engagement would be useless—thus resorting to very expensive f2f meetings. The findings here, however, indicate that online discussions can be as useful in undermining stealth democracy and related beliefs as f2f discussions. In particular, the only condition that consistently and significantly ameliorates all three stealth democracy and related beliefs here is the online discussion condition with no reminders of citizenship. This condition appreciably reduces vertical collectivism and perceived conflict, as well as stealth beliefs.

To closely replicate the condition here that consistently undermined stealth democracy beliefs and some of its problematic attitudinal precursors, digital government practitioners should set up audio-based online discussions using full participant names, avoid use of symbols of citizenship (flags, the word "citizen", references to the country), and appeal to participants to take the occasion to learn about what is useful to make up their own minds—avoiding mention of their role as community members. No text-based deliberations were tested here, so it is possible that audio is unnecessary for the desired effects. Indeed, text-based discussion may have stronger effects. Being online consistently matters only for reducing vertical collectivism (VC). Those online with citizen reminders may be so absorbed in the citizen role that they do not experience conflict between their own wishes and that of the group, also reducing VC. Those online with individuality reminders may become more individualistic and reject the needs of the group, undermining VC. If the latter explanation is correct, then text-based discussion, which is even more anonymous than audio, should enhance the reduction of VC because it creates greater deindividuation.

To achieve the positive effects described here, government officials need to introduce online deliberation in such enterprises as e-rulemaking. Rulemaking is among the most prominent ways in which citizens can provide input into government [29]. Current e-rulemaking systems discourage discussion by participants by structuring input as individual comment documents rather than as discussion threads. Public comments typically flood in toward the very end of the assigned discussion period for a rule, limiting

interaction. Also, public interest groups have approached rulemaking as a plebiscite by taking public comments as an opportunity to flood officials with form letters, precisely not what federal officials find helpful [29]. In future work, I hope to test several methods of online deliberation in actual e-rulemaking. Numerous freeware products already exist for real-time chat or threaded bulletin board discussions such as PHP Website, mvnForum, GroupServer, phpBB, and Deme. One particularly exciting possibility includes using Second Life, a virtual reality environment with text chat and possibilities for audio, to host online e-rulemaking deliberations. An interesting question is whether having an avatar will prove more or less deindividuating than standard textual communications.

8.ACKNOWLEDGMENTS

This research is based upon work supported by the National Science Foundation under Grant No. EIA-0205502.

9.REFERENCES

- [1] G. S. Aikens, "A History of Minnesota Electronic Democracy 1994," *First Monday*, vol. 1, pp. <http://www.firstmonday.dk/issues/issue5/aikens/#dep3>, 1996.
- [2] B. Altemeyer, *The Authoritarian Specter*. Cambridge, MA : Harvard University Press, 1996.
- [3] J. T. Cacioppo, R. E. Petty, J. A. Feinstein, and W. B. G. Jarvis, "Dispositional Differences in Cognitive Motivation: The Life and Times of Individuals Varying in Need for Cognition," *Psychological Bulletin*, vol. 119, pp. 197-253, 1996.
- [4] R. Carltz and R. Gunn, "e-Rulemaking: a New Avenue for Public Engagement," *Journal of Public Deliberation*, vol. 1, 2005.
- [5] P. E. Converse, "The Nature of Belief Systems in Mass Publics," in *Ideology and Discontent*, D. E. Apter, Ed.: Free Press, 1964, pp. 206-261.
- [6] M. H. Davis, "Measuring Individual Differences in Empathy: Evidence for a Multidimensional Approach," *Journal of Personality and Social Psychology*, vol. 44, 1983.
- [7] M. X. Delli Carpini and S. Keeter, *What Americans Know About Politics and Why it Matters*. New Haven, Conn.: Yale University Press, 1996.
- [8] J. S. Fishkin, *The Voice of the People: Public Opinion and Democracy*. New Haven, CT: Yale University Press, 1997.
- [9] J. R. Hibbing and E. Theiss-Morse, *Stealth Democracy: Americans' Beliefs About How Government Should Work*. Cambridge, U.K. ; New York: Cambridge University Press, 2002.
- [10] D. R. Kinder, "Diversity and Complexity in American Public Opinion," in *Political Science, The State of the Discipline*, A. Finifter, Ed. Washington: American Political Science Association, 1983, pp. 391-401.
- [11] B. Kirschner, "PEN Lessons: An Interview with Ken Phillips," *Public management*, pp. 13, 1994.
- [12] R. C. Luskin, J. S. Fishkin, and R. Jowell, "Considered Opinions: Deliberative Polling in Britain," *British Journal of Political Science*, vol. 32, pp. 455-488, 2002.
- [13] M. A. Milburn, S. D. Conrad, F. Sala, and S. Carberry, "Childhood Punishment, Denial, and Political Attitudes," *Political Psychology*, vol. 16, pp. 447-478, 1995.
- [14] P. Muhlberger, "Human Agency and the Revitalization of the Public Sphere," *Political Communication*, vol. 22, pp. 163-178, 2005.
- [15] S. L. Neuberg and J. T. Newson, "Personal Need for Structure: Individual Differences in the Desire for Simpler Structure," *Journal of Personality and Social Psychology*, vol. 65, pp. 113-131, 1993.
- [16] W. R. Neuman, *The Paradox of Mass Politics: Knowledge and Opinion in the American Electorate*. Cambridge, Mass.: Harvard University Press, 1986.
- [17] B. S. Noveck, "The Electronic Revolution in Rulemaking (<http://ssrn.com/abstract=506662>)," *Emory Law Journal*, 2004.
- [18] T. Postmes, R. Spears, and M. Lea, "Breaching or building social boundaries? SIDE-effects of computer-mediated communication," *Communication Research*, vol. 25, pp. 689-715, 1998.
- [19] V. Price and J. N. Cappella, "Online Deliberation and its Influence: The Electronic Dialogue Project in Campaign 2000," *IT&Society*, vol. 1, pp. 303-329, 2002.
- [20] V. Price and C. David, "Talking About Elections: A Study of Patterns in Citizen Deliberation Online," presented at International Communication Association Annual Meeting, New York, NY, 2005.
- [21] R. D. Putnam, *Bowling Alone: The Collapse and Revival of American Community*. New York: Simon & Schuster, 2000.
- [22] A. E. Raftery, "Bayesian Model Selection in Social Research," *Sociological Methodology*, vol. 25, pp. 111-163, 1995.
- [23] H. Rheingold, *The Virtual Community: Homesteading on the Electronic Frontier*, Rev. , 1st MIT Press ed. Cambridge, Mass.: MIT Press, 2000.
- [24] S. W. Rosenberg, *The Not So Common Sense: Differences in How People Judge Social and Political Life*. New Haven, CT: Yale University Press, 2002.
- [25] S. W. Rosenberg, *Reason, Ideology and Politics*. Princeton, New Jersey: Princeton University Press, 1988.
- [26] L. Ross and A. Ward, "Naive Realism in Everyday Life: Implications for Social Conflict and Misunderstanding," in *Values and Knowledge*, E. S. Reed and E. Turiel, Eds. Mahwah, NJ: Lawrence Erlbaum Associates, 1996, pp. 103-135.
- [27] E. A. Schwartz, *Netactivism: How Citizens Use the Internet*, 1st ed. Sebastopol, CA: Songline Studios, Inc., 1996.
- [28] R. Shapiro, "Stealth Democracy: Americans' Beliefs about How Government Should Work by John R Hibbing and Elizabeth Theiss-Morse," *Political Science Quarterly*, vol. 118, pp. 2, 2003.
- [29] S. W. Shulman, "The Internet Still Might (but Probably Won't) Change Everything: Stakeholder Views on the Future of Electronic Rulemaking," University of Pittsburgh / NSF, Pittsburgh 2004.
- [30] S. W. Shulman, D. Schlosberg, S. Zavestoski, and D.

- Courard-Hauri, "Electronic Rulemaking: New Frontiers in Public Participation," *Social Science Computer Review*, vol. 21, pp. 162-178, 2003.
- [31] J. Sidanius and F. Pratto, *Social dominance : an intergroup theory of social hierarchy and oppression*. Cambridge, UK ; New York: Cambridge University Press, 1999.
- [32] R. Spears, M. Lea, and S. Lee, "De-individuation and group polarization in computer-mediated communication," *British Journal of Social Psychology*, vol. 29, pp. 121-134, 1990.
- [33] H. C. Triandis, "The Psychological Measurement of Cultural Syndromes," *American Psychologist*, vol. 51, pp. 407-415, 1996.

Converting Online Public Legal Information into Knowledge : “ABC del Diritto” an Italian e-Government Citizen-oriented Service*

Maria Angela Biasiotti
Istituto di Teoria e Tecniche dell'Informazione Giuridica
via de' Barucci 20
Florence, 50127, Italy
biasiotti@ittig.cnr.it

Roberta Nannucci
Istituto di Teoria e Tecniche dell'Informazione Giuridica
via de' Barucci 20
Florence, 50127, Italy
nannucci@ittig.cnr.it

ABSTRACT

Keywords

e-Government Services to Citizens, Public Administration, Online Legal Information

1. INTRODUCTION

e-Government is the use of information and communication technologies in public administration - combined with organisational change and new skills - to improve public services and democratic processes and to strengthen support to public policies. e-Government is a way for public administration to become more open and transparent, and to reinforce democratic participation; more service-oriented, providing personalised and inclusive services to each citizen.

e-Government implies both ICTs and human resources: whereas governments are suppliers of the e-Government system, end-users/citizens are its customers. So, the implementation of *e-Government*, while implying the modernization of procedures and structures within PA organizations (that is *e-Administration*), regards also the change of procedures and modalities in which citizens and PA relate each other (that is *e-Democracy*), and all together aim at achieving a new way of ruling public matters, that is a new *Governance* or *e-Governance*. Therefore, *e-Government* with all its strategic outlined roles becomes an essential step in the development process of a new governance form as hoped and promoted by the European Union, whereas it involves not only the improvement of services quality, but mainly makes users more aware of public activities in progress and favours their active participation as citizens.

- - -

Generally speaking the specific goals of e-Government services should be to enhance citizens awareness and to convert available information into achievable knowledge.

The understanding of people who will be using e-Government services is therefore critical for creating good added-value services.

This should extend to the way they understand computers and the internet, the ways in which they think about and carry out tasks, and the context in which they will do this. Understanding, however, is one thing, an effective use is something else. Even if a service might be created to be useful to citizens but does not take into account how it can be utilized by them, the constraints on their attention, their level of technical ability, the difficulty of accessing it and finding relevant data, then it is almost unlikely that it will be used at all. This type of service is of little interest for citizens.

Presently ABC for the law citizen/user target is mainly represented by those categories able to retrieve online information and to use possibilities offered by Information Technologies tecniques in a proper way. This happens with those persons who have a sufficient digital literacy background. In the Italian panorama this is true for some categories of young professionals such as academics, lawyers, accounters, administrators and not really for the common citizen. Nevertheless, as digital literacy is increansingly being diffused to the general public, access to online services such as ABC are growning as it is possible to verify by system monitoring. Finally, also the Public Sector Information Directive Directive 2003/98/EC on the re-use of public sector information adopted by the European Parliament and by the Council on 17 November 2003 was designed to make it easier for content producers to use and add value to information produced by the public sector, both providing useful content for the development of the Information Society and making public sector content more accessible to more people.

1.1 Objectives - Benefits from e-Government Citizen-oriented Services: from Information to Knowledge

This paper aims at demonstrating that citizen-oriented services such as ABC for the Law are able to facilitate access of citizens to the legal information offered by public administrations converting information into real knowledge using the possibilities arising from ICTs. Specifically the methodology applied for building up the system focuses the attention on the identification of those legal concepts involved in the legislation on the net search and therefore to be acquired by the user in order to understand the result. So that ABC can also be considered a sort of tutorial guiding the user in his searching and making him understand the data he finds in a more conscious way. Offering additional information about other related sources available online and widens the user research strategies and increases his knowledge background. This is also possible as ABC considers all legal concepts involved when searching for legislation on the net.

1.1.1 Information Dissemination State of the Art

Through cyberspace people are presently involved in ways never envisioned before: they are often overwhelmed by information which is frequently not exhaustively clear as to contents and language. For these reasons and for other technical barriers it is evident that information itself is not yet knowledge. Citizens need places where information can be transformed into knowledge, that is shared understanding. Specifically, recent research into e-Government practices and applications illustrates that the creation of specific tools for sharing and developing real knowledge is necessary for enhancing citizens involvement at least in this early stage of e-Government implementations, when the access to a large spectrum of information is simply preferred to a concrete communication between citizens and PAs. Information may become knowledge only when great attention is given to key elements such as contents and actors of the communication process and practical solutions are identified and elaborated in relation to the specific field of interest (Law, Economy, Social, Health etc). Reaching these results requires practical groundwork, starting small with innovations, learning from experience, sharing of best practice and finding scalable solutions. As to legal knowledge the dissemination of official public information and documents is becoming increasingly important as they are the very nature pre-eminent examples of public sector information, which relies mostly on universal access, in the sense that access should be freely available. Nevertheless, at present governments are not yet sufficiently open and transparent to induce citizens involvement. Information is often represented in the net in a way which is unclearly organized both for expert and non expert users, as the majority of citizens are. Furthermore dealing with public information and data means to be in touch with qualified sources mostly regarding domains not easily comprehensible by common users. With the aim to mediate between PA features and citizens' awareness it seems therefore very adequate to adopt specialized knowledge-oriented support tools, such as information system guidelines, structured search engines, guided navigation paths, specialized glossaries, user-oriented illustrative and recapitulatory tables for specific domains such as law, economy and other disciplines which may contribute to enhance citizens' capability to access online

information. At present there are many experimental applications carried out in several European countries and much discussion is held on which are the best means for transforming information into knowledge, and which structure and format they may have [4], [1], [5]. Another example can be considered Informiran.si [8], a system implemented in Slovenia aiming at aiding common users such as local citizens to fill legal documents online by means of a sort of electronic guide of relevant information necessary for achieving these goals.

1.1.2 ABC for the Law Framework

An important step of Italian e-Government strategies is represented by specific actions for enhancing the communication between government and citizens. It is in this framework that ABC for the Law was built as a tool supporting citizens when accessing and consulting the *NormeInRete* portal. NormeInRete (Legislation on the Net) Project, promoted by the Italian Authority for Information Technology in the Public Administration (AIPA) and the Ministry of Justice in collaboration with the Institute of Legal Information Theory and Techniques of the Italian National Research Council (ITIIG/CNR), aims at fulfilling the citizen's right to acquire knowledge about legislation and supports the Public Administration in managing the legislative documentation life cycle efficiently. More specifically, the *NormeInRete* Project (NiR) aims at improving accessibility to legislation by providing a unique access to Italian and European Union legal documents published on different websites through a specialised portal (www.nir.it).

The NormeInRete portal runs a search engine that operates uniformly on distributed data sources. Its full text search index is selectively built to detect only legislative documents. The achievement of a higher level of co-operation relies on the adoption of two standards, defined within the Project by ad hoc Working Groups in which major PAs and research institutions have taken part. The standards have been issued as AIPA technical standards and published as regulations in the Italian Official Journal. The definitions make use of Uniform Resource Names (URNs) (RFC 2141) and eXtensible Markup Language (XML W3C Recommendation) standards.

ABC for the Law can be considered an example of a knowledge support tool through which citizens can access information in a more thorough way and at the same time learn about specific concepts or topics studying them deeply by using abilities incorporated in the system.

2. METHODOLOGY AND CASE DESCRIPTION - MANAGING LEGAL KNOWLEDGE IN E-GOVERNMENT CITIZEN-ORIENTED SERVICES: "ABC FOR THE LAW"

Consulting and accessing legislation implies for citizens a basic legal background for orienting their search and needs. Citizens are not able to search for a Government decree if they do not know the difference between this kind of document and a law of the Parliament; moreover, they cannot search for a law provision if they do not know that a law be-

Ricerca le norme

Estremi del provvedimento:

tipo: nessuna selezione

numero: _____ anno: nessuno

Cerca nel testo:
contiene le parole:

Tutte le parole Almeno una parola Frase esatta

distanza massima: indefinita

Cerca Reimposta

Figure 1: NIR Search Interface

fore being valid should be approved and, for the Italian system, edited in the Official Gazette. Furthermore, in the European context it is important for citizens to understand the relationship between the European Union and its member States in order to become aware of their European rights and duties. In order to expand the access not only to experts but also to common users it was considered essential to support the legislation search with a specifically built less complex tool facilitating the comprehension of legal basic concepts and guiding users towards their specific goals. Taking into account recent research and surveys on stakeholders accessing online public information and their peculiarities (such as educational background, familiarity with legal concepts and capability to navigating in the net) two aspects were mainly to be faced for enlarging access to general public: the choice of relevant concepts and the way these were to be presented.

*ABC for the Law*¹, is a specific tool recently elaborated in 2004 by ITTIG/CNR, the Ministry of Justice - Direzione generale per i Sistemi Informativi Automatizzati (DGSIA) and Centra Nazionale per l'Informatica nella Pubblica Amministrazione (CNIPA) with the aim of explaining some basic legal concepts, the knowledge of which seems useful for supporting citizens in their legislation search. When the Legislation on the Net portal was initially developed, it was mainly addressed to expert professionals such as lawyers, public officers and judges as the query forms were simply based on technical legislation references - such as law typology, official identification number, issue year and specific domain keywords - implying as such a thorough and deep knowledge of the Law (see Figure 1).

Particularly, ABC contents was chosen with reference to the nature and purposes of the Legislation on the Net portal, that is access to public information requiring a good level of legal background. Basic legal concepts were identified and introduced for guaranteeing an elementary acquaintance of Constitutional law, Public law, Civil law, Criminal law, European and International law. The language employed in ABC was simplified and turned to render easily comprehensible the legal concepts included in the laws to be accessed as if the accessing citizens were represented by students of secondary schools. In fact a first draft of *ABC for the Law* was distributed to a class of a technical secondary school where Law is taught as part of the curriculum and their feedback was taken into account. The system content is subdivided into 8 parts, and these are articulated into other subsessions.

¹<http://www.normeinrete.it/abc/html/indice.htm>



Figure 2: ABC Homepage

As indicated in the title page left side, the main parts are: Law definitions and classification; Law Sources; Italian Law sources; European Law sources; International Law sources; Italian Legal system; European Union Institutions; International organisations (see Figure 2). As to the way in which concepts are introduced, the attempt was made to indicate search paths to facilitate users' involvement in transforming simple information into effective knowledge. The covered arguments, as indicated in the right side, are enriched with numerous hypertextual links of internal nature, identified with an [I], and of external nature, identified with an [E]. The former are conceived for enabling the user to deepen the knowledge he is searching for allowing him to move easily within the entire text and also to read through some relevant and more specific information or data; the latter were conceived to allow citizens to connect directly to the sources available in the net (i.e. institutional websites) for better understanding the concepts under consideration and for updating and contextualizing searched information.

The framework of the internal links is also enriched by some references identified by [T], that address the user to some illustrative tables on particularly complex topics or procedures, which represent an instrument of immediate perception and a key for a synthetic and simplified reading and comprehension (see Figure 3).



Figure 3: ABC Html Page

A specialized legal Glossary allowing the consultation of short definitions of relevant or difficult terms and concepts dealing with the ABC domain has been added and may be consulted independently from the consultation of the complete system or by accessing the specific link identified with a [G] within the ABC contents. From a technical point of view ABC was implemented by means of Macromedia Dreamweaver MX; pages are static but their layout is man-

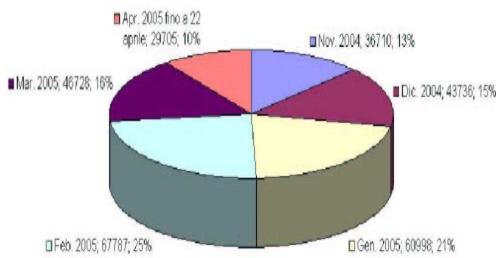


Figure 4: ABC Accessing Report (November 2004-April 2005)

aged by templates according to the features of the hosting system. From a first statistical analysis monitored by the Ministry of Justice on *ABC for the Law* after its recent insertion into the NIR portal (last six months), some issues come up which may deserve some attention. The *ABC for the Law* system was accessed by a great number of users (about 280.000), especially soon after its electronic publication. As it can be seen in the table represented in Figure 4 access was concentrated in January and February.

From a more detailed analysis of these accesses (based on automated lists of accesses to single pages by log files), the 15% of users navigated into *ABC* html pages concentrating especially on its two first sections (Law definitions and classification; Law Sources), on the illustrative tables and on the Glossary: it demonstrating - even if with small numbers - that citizens, when searching for legislation on the net, appreciated the support of simple explanations of too theoretical concepts as well as the aid of graphical and synthetical representations of complex procedures.

3. CONCLUSIONS AND RECOMMENDATIONS

These first statistical results deriving from a brief monitoring of *ABC for the Law* accesses seem to show that the integration of online information systems with citizen-oriented support tools, specifically in legal domains significant in the relationship between citizens and public administration, is appreciated by users and capable of enriching their capacities. It seems therefore worthwhile to develop and adopt support tools capable of promoting a more active citizens involvement. Access to online information/knowledge is currently difficult. Two aspects are to be underlined: from one side the language used is very often too technical and not oriented to non-expert citizens; from the other side the huge amount of data and information on the net - especially in institutional websites, implemented by public administration - are not transparent and this does not help citizens in their access to information and search of documents. In both cases the adoption of a support tool such as ABC may become a key challenge for Governments to organize, classify and manage information in a more rational way, and for citizens to be oriented in their navigation according to them specific needs, as, starting from a certain concept or word they may choose (as in ABC), they may be guided within the website contents through useful links and connections. At the moment ABC for the Law is running in

the NormeInRete portal as an additional service for citizens accessing the legislation on the Net portal. We are in the second implementation phase where a large survey among different categories of ABC stakeholders is being done in order to assess ABC for the Law user satisfaction and to test the efficiency, usability (ease of use), utility, accessibility and quality of the service itself [7]. Literature suggests many different types of usability evaluation methods basically divisible by three categories: inspection, inquiry and formal usability testing. Inspection is mainly based on experts evaluating and examining usability related aspect of the service; inquiries involve users concretely experienced to evaluate their preferences, experiences and expectations with a site; formal usability testing can be conducted as an experiment in real situations [6]. The goal is to validate the system and the user satisfaction using all these usability evaluation methods. Further improvements can be already proposed towards this prospect such as the development and introduction into the system of proper metadata and artificial intelligence features. A significant step should be the creation of a specific ontology for structuring the complete legal content of ABC, facilitating in this way the logical approach to the system. Aids such as *ABC for the Law* represent a valid means for converting information into knowledge and for educating citizens both in the use of technological tools and in their acquaintance of a basic legal approach favouring the awareness of their rights and of their possible actions in order to take active part to present and future social transformations. Within this framework it might be considered as a starting conceptual core to be submitted to an e-Learning platform in order to build a system which, after due experimentation, might become a tool far more able to induce knowledge; this would increase the application of the system to citizens lifelong learning and to public officers' updating. Furthermore, the development of these types of tools represents also an innovative approach within the framework of European strategies aiming at building the Information Society. One major priority of the eEurope 2005 Action Plan is that all Europeans must have the opportunity to develop the skills necessary to participate in the Information Society and take advantage of the range of technologies and services available. Although many efforts have been devoted by national governments to digitalizing contents and services, there is still a long way from achieving those goals related most closely to social inclusion and the knowledge-based society. It is up to governments to further promote and elaborate proper services and tools; it is up to citizens to exploit implemented technological opportunities [2], [3].

4. REFERENCES

- [1] B. L. e. B.N. Hague. *Digital Democracy: Discourse and Decision-making in the Information Age*. Routledge, New York, 2002.
- [2] D. Boulle. Reinventing democratic culture in an age of electronic networks.
<http://www.netaction.org/boulle/index.html>.
- [3] M. Castells. *The Rise of the Network Society*. Blackwell, Oxford, 2000.
- [4] S. Clift. Democracy is online. *A Journal of the Internet Society*, March/April 1998.

- [5] O. (ed.). Citizens as partners. information, consultation and public participation in policy making.
<http://www1.oecd.org/publications/e-book/4201131e.pdf>.
- [6] A. T. M. Waiguny. Usability evaluation fro reletionship-oriented web sites. In *Paul Cunningham and Miriam Cunningham, Innovation and the Knowledge Economy. Issues, Applications, Case Studies.* IOS Press, The Netherlands, 2005.
- [7] I. Milicevic, G. Karsten, and W. Korte. Making progress towards user-orientation in online public service provision in europe. In *Paul Cunningham and Miriam Cunningham, Innovation and the Knowledge Economy. Issues, Applications, Case Studies.* IOS Press, The Netherlands, 2005.
- [8] A. Tomazic, M. Jamnik, and P. Licen. Environment and tools for creating legal documents online. In *M. Palmirani, T. van Engers, R. Traunmueller - The Role of Knowledge in e-Government.* AAIAL, 2005.

Web Portal Implementation to Support Public Participation in Transportation Decision Making

Tim Nyerges

University of Washington

Dept of Geography, Box 353550

Seattle WA 98195

1.206.543.5296

nyerges@u.washington.edu

Terry Brooks

University of Washington

Information School, Box 352840

Seattle WA 98195

1.206.543.2646

tabrooks@u.washington.edu

Piotr Jankowski

San Diego State University

Dept of Geog, 5500 Campanile Drive

San Diego CA 92182-4493

1.619.594.0640

piotr@geography.sdsu.edu

G. Scott Rutherford

University of Washington

Dept of Civil & Envir Engr, Box 352130

Seattle, WA 98195

1.206.685.2481

scottrut@u.washington.edu

Rhonda Young

University of Wyoming

Dept of Civil & Arch Engr, Dept 3295

Larame, WY 82071-3245

1 307 766 2184

rkyoung@uwyo.edu

ABSTRACT

In this paper we describe the start of system implementation in the participatory geographic information system for transportation (PGIST) project. The PGIST web portal is being developed to support public participation in transportation improvement decision making in the central Puget Sound region of Washington State. Implementation has followed design considerations. A web services architecture forms the basis of the implementation.

Categories and Subject Descriptors

H.4.2 [Information Systems Applications]: Decision support systems – *public participation geographic information systems, collaborative computing*.

General Terms

Human Factors, Decision Support, Portal Development.

Keywords

Geographic information systems, public participation, group decision making, decision support.

1. INTRODUCTION

Research about local governance involving public-oriented (e.g., environmental, transportation, land use) decision making suggests little ‘meaningful public participation’ exists. Meaningful public participation is about *access to voice* and *competence of knowledge(s)* that foster *shared understanding* about valued-concerns. Research about analytic-deliberative decision processes shows that meaningful public participation is possible, and decision outcomes are improved in relation to such concerns. Although analytic-deliberative decision settings have been convened in synchronous settings, particularly in small groups, there have been no settings that support large-groups (e.g., 100 or more people) in analytic-deliberative public participation. To fill this gap, we are developing asynchronous, Internet-based geographic information system (GIS) portal to support creation and use of analytic information structures (e.g., maps, tables,

models) together with on-line deliberative discourse. The application setting is regional transportation improvement program (TIP) decision making. The principal research question is: What Internet portal designs and capabilities, particularly including geographic information system (GIS) technology, can enable public participation in analytic-deliberative transportation improvement decision making within large groups?

The portal design is based on a web services architecture using five functional areas: Agenda Management, Concerns-Values Organization, Alternatives Generation, Choice Modeling, and Reflective Review. The Agenda Manager establishes decision situation and meeting agendas. The Concerns-Values Organizer builds stakeholder concerns hierarchies about transportation improvement. The Alternative Generator builds scenarios composed of collections of transportation projects characterized in terms of social, economic and environmental criteria. The Choice Modeler supports trade-off analysis of scenario impacts. The Reflective Review is a feedback mechanism available at all times so users can comment on any aspect of the capabilities.

Each of the functional areas is a collection of analytic-deliberative tools supporting participatory modeling across a TIP decision process. We have used an interaction design approach, focusing on multiple levels of human-computer-human interaction process.

2. RECENT PROGRESS

Project progress focuses on software design and implementation, database development, and usability evaluation.

2.1 Software Design and Implementation

We have developed a participatory activities framework (PAF) for characterizing human-computer-human interaction (HCHI) at six levels of process granularity. The PAF levels form an aggregation (building block like) description of analytic-deliberative HCHI ranging from speech acts at the shortest process level (top row in Table 1) all the way to the TIP decision situation (bottom row in Table 1). The event process time runs from seconds for data operations (e.g., speech acts) to months for

the decision situation. The PAF describes the granularity of portal functionality as an embedded hierarchy from level to level, providing a foundation for software design. The PAF separates analytic and deliberative processes at the shortest three of the six levels of process resolution. However, there is considerable interaction between the two types of processes. Although the idea of separating analysis from deliberation is really a false binary, it is necessary to design and implement analytic and deliberative functionality.

Table 1. Participatory Activities Framework

Analytic Activity Focus	Deliberative Activity Focus
<i>Analytic data act</i> – an analytic data operation.	<i>Deliberative data act</i> – a deliberative data operation.
<i>Analytic information act</i> (AIA) formed from sequencing analytic data acts; considered a participatory act (P-act) in a group setting.	<i>Deliberative information act</i> (DIA) formed from sequencing speech data acts; considered a participatory act (P-act) in a group setting.
<i>An-game</i> – Analytic game is composed of a series of AIAs, and is called a <i>P-game</i> within a group setting.	<i>De-game</i> – Deliberative game is a series of DIAs, and is called a <i>P-game</i> within a group setting.
<i>Participatory method</i> is a series of P-games that structures analytic-deliberative groupwork process	
<i>Decision support meeting session</i> is a series of one or more participatory methods	
<i>Decision situation</i> is composed of one or more decision support meeting sessions	

The activities of the PAF provide a sense of the granularity of analytic and deliberative objects that embed and interact. The embedding occurs from row to row down the table within each column. Hence, data acts compose information acts. Information acts combine to form the basis of participatory games. Information acts are the basis of operations for information structures, (e.g., maps, models, tables) that get put to use in p-games (e.g., brainstorming, idea synthesis, voting activities). P-games are sequenced as a series of interactivity. P-games have “rules of engagement” for helping people understand the protocols for interacting. P-games are the fundamental steps of structured participation methods, for example, Delphi, Nominal Group, Technology of Participation, and Citizen Jury. Because of a variety of process descriptions, there are micro, meso, and macro P-games that can be used to compose the steps of P-methods, but we have yet to enumerate all of them. P-methods support effective and equitable analytic-deliberative expressions to elicit concerns about transportation improvement, relate those concerns to scenario and project-based alternatives, then allow those alternatives to be compared using sensitivity trade-offs that underpin interactive equity analyses. Drawing together multiple levels of process granularity from diverse literatures provides us with new theoretical insight into HCHI. Eventually, this new understanding will allow designers to compose libraries of analytic-deliberative process techniques to enable large numbers of people to engage in public participation.

Implementation of these capabilities is proceeding based on function screen mockups that incorporate the six levels of participatory activity as appropriate. Those mockups represent the design level artifacts of the system development effort. A web

services architecture is being used to organize the implementation of prototypes. The Concerns-Valued Organizer based on natural language processing continues to be our biggest challenge for system development. Unpacking and integrating the “meaning” of speech acts in discourse is the core of that challenge.

We are using Servlet technology and Struts web application framework. We are coding content objects using XML and AJAX for client-side scripting. The web application architecture is hosted in a Tomcat 5.x web server with PostGreSQL 8.0 database management and MS Windows Server 2000.

We are managing the development effort using concurrent versioning system (CVS). The top-to-bottom management of artifacts has been a challenge, but we have managed to make the best use of the Plone content manager for this purpose.

2.2 TIP Project Database Development

The second major aspect of our development effort is the database focusing on major regional TIP projects. The database development has been proceeding on three fronts. First, there is a requirement to implement the TIP projects in regards to their spatial-temporal footprint. Second, those projects are described in terms of the agency (institutional) considerations that appear in a variety of planning and improvement programming documents, and expressed in terms of objectives and criteria for improvement. The data elements for those characteristics are being obtained from multiple public agencies. Third, transportation improvement fosters public concerns about enhancing various aspects of regional and local community. These data are to be obtained during specialist panels for review by the public.

2.3 Usability Evaluation

As we move forward with system implementation, we are formulating alpha and beta-level usability tests. We are using the decision scenarios and personas developed for system design. Our decision scenarios will form the basis of the default decision processes for the decision experiment we call “Let’s Improvement Transportation” (LIT). Personas are essentially *user types* that researchers adopt to exercise the system in those scenarios. The scenarios and personas set the stage for usability tests of system characteristics in anticipation of the LIT decision experiment.

3. CONCLUSIONS

Software, database, and usability implementation are based on a mix of information needs elicited from partners at the beginning of the project, together with case studies of previous decision situations with the same partners, plus ideal public participation scenarios reported in research, and the follow up designs for the portal. Implementation remains as complicated as the design, but current findings show some very interesting conceptual developments well worth the effort. Publications and progress updates can be found at www.pgist.org.

4. ACKNOWLEDGMENTS

This research has been supported in part by National Science Foundation Grant No. EIA 0325916, funded through the Information Technology Research Program, and managed in the Digital Government Program. The authors are solely responsible for the content.

SESSION 3A

CRISIS MANAGEMENT 1

Moderator

José Fortes, University of Florida, USA

Titles and Authors

GeoCollaborative Crisis Management: Designing Technologies to Meet Real-World Needs
MacEachren, Alan M.; Cai, Guoray; McNeese, Michael; Sharma, Rajeev; Fuhrmann, Sven

Indexing and Searching Handwritten Medical Forms
Govindaraju, Venu

Digital Governance and Hotspot GeoInformatics for Monitoring, Etiology, Early Warning, and Management Around the World
Patil, G.P.

GeoCollaborative Crisis Management: Designing Technologies to Meet Real-World Needs

Alan M. MacEachren ¹⁺², Guoray Cai ¹⁺³, Michael McNeese ¹⁺³, Rajeev Sharma ^{1,4+5}, Sven Fuhrmann ⁶

¹ GeoVISTA Center, Penn State University, University Park, PA 16802

² Department of Geography, Penn State University, University Park, PA 16802

³ College of Information Sciences & Technology, Penn State University, University Park, PA 16802

⁴ Dept. of Computer Science & Engineering, Penn State University, University Park, PA 16802

⁵ VideoMining, Inc. State College PA

⁵ Dept. of Geography, Texas State University, San Marcos, TX

{maceachren, gxc26, mdm25, rxs51}@psu.edu

ABSTRACT

Preventing, preparing for, responding to, and recovering from natural and human-induced disasters all require access to geographically referenced information and tools for making available information relevant to the tasks at hand. Goals of the research summarized here are to advance our scientific understanding of how groups (or groups of groups) work with geospatial information and technologies in crisis management and to use that understanding to guide development of tools that are intuitive for non-specialist users and that enable coordination within and across crisis management teams. This overview highlights progress on: understanding work in crisis management, enabling distributed information access through context-mediated geo-semantic interoperability, extension of natural, multimodal interface methods to mobile devices, development of a collaborative map-based web portal to support international humanitarian relief logistics, and technology transition into real-world practice. We also introduce our new DHS-supported Regional Visualization & Analytics Center, which builds directly upon our GCCM work.

Keywords

Multimodal Interfaces, Knowledge Elicitation, Human-Centered Design, GeoCollaboration, GIS, Crisis Management.

1. INTRODUCTION

The challenges faced by government and other organizations charged with responsibility for crisis management is immense. Information technology is fundamental in efforts to assess vulnerabilities, prevent undesirable events, minimize event impacts, and enable rapid recovery. Geographically referenced information and supporting technologies are central to many crisis management tasks and have been used effectively at all scales (from local ambulance dispatch, through response to major terrorist actions, to regional scale response such as that for hurricane Katrina).

Current geospatial information technologies (GITs) create the potential to integrate diverse information quickly; however, the

technologies remain hard to use and ill-suited to group work. Impediments to wide-spread, real-world use include overly complex interfaces, limited support for analytical reasoning and decision-making, and for coordinated team work. Underlying these functional limitations are major gaps in understanding of how these technologies facilitate or impede individual and group work.

As outlined previously [1], our research addresses two overarching issues: (1) the understanding of GIT-based individual and group work in crisis situations and (2) the development of GIT that enables coordinated same-place and distributed crisis management activities. More specifically, our focus is on: (a) understanding cognitive readiness in real world geo-collaborative activity; (b) testing theories of cognitive readiness within team simulation environments; (c) understanding technology enabled group work; (d) developing natural, easy to use, multimodal interfaces to geospatial information technology, and (e) developing Computer Supported Collaborative Work Systems (CSCW) that use shared visual displays to mediate discussion of site situation, and action for crisis management.

2. ACCOMPLISHMENTS

During the past year, progress has been made on fundamental science questions as well as on technology implementation and its transition to real world applications. Technology transition will be touched upon in the next section. Here we sketch key science and technology implementation accomplishments in four areas.

2.1 Understanding work

We have applied the Living Lab Framework [2] as an integrated approach to understanding work in the real world and modeling that work to support testing of theory and technology. The approach includes: (1) *Cognitive fieldwork* within several crisis management domains (911 centers, police operations, emergency medical applications, and terrorist training exercises) [3]; (2) *Simulations* (using NeoCITIES, see: [4]) based upon realistic scenarios for emergencies that demand that team cognitive processes operate in uncertain, ill-defined situations. This provides a basis to develop user interfaces and visual analytics methods that address constraints present in teamwork and technology and produce more effective and efficient support of cognitive readiness.

2.2 Context and geo-semantic interoperability

Geospatial data semantics deal with representations of the geographical world as interpreted by individual human users or a community of practitioners. Ontology-based approaches have been accepted as the panacea for all sorts of geospatial semantic

problems, and identifying categories, concepts, relations, and rules that prescribe theories of the geospatial domain.

Our work proposes an extension of ontology-based methods with an explicit model of context that broadly characterizes typical applications and scenarios of use, and complements traditional abstraction and modeling methods. Specifically, we propose a semantic model of geographical data that supports the reasoning of spatial data meaning based on context. Conceptually, we represent geospatial database semantic knowledge using a contextualized geo-ontology that represents an unambiguous and coherent theory about a piece of geographical reality within a prescribed context. Multi-range contexts allow multiple ontologies to co-exist in a system and jointly describe multiple data source semantics. At run-time, data sharing or communication contexts are used to mediate ontology alignment and semantic conflict resolution. This is accomplished through an intelligent agent that explicitly captures knowledge about defining features and proper behaviors of a context in the form of contextual schemas (C-schemas) [5], and provides a holistic solution to geospatial semantics.

2.3 Adaptive interfaces for mobile devices

The dynamic nature of user multitasking, work environment switching, and shifts of internal goals make context awareness, relevancy and adaptation critical in mobile computing. Existing studies and models of mobile contexts recognize physical, cognitive and social contexts as major categories, but are unclear about contextual influence on mobile application behavior and treat context using ad hoc adaptation strategies. We address these issues by establishing a computational model of mobile context relevancy [6] that binds contexts, activities and adaptation strategies at run-time with a degree of consciousness about changing contexts and an autonomy in choosing a proper dynamic adaptation strategy. This system actively regenerates a model of ongoing activity by sensing, communicating, and interpreting changing conditions, resources and processes. This allows contextual factors to be associated with an activity based on how factors contribute to various components of an evolving collaborative plan.

2.4 Collaborative map-based portal

A Geocollaborative Web Portal (GWP) application has been designed and (partially) implemented to provide a common interface that supports asynchronous, geocollaborative activities for humanitarian relief logistics operations. The GWP emphasizes support for situation assessment, positioning and monitoring of field-teams and distribution sites, and supply routing. For details, see our demonstration description (Tomaszewski, et al, this volume).

3. TWO SHORT SUCCESS STORIES

3.1 Port Authority of NY and NJ (PANYNJ)

In June 2005, an alpha version of the Geospatial Multimodal Interaction Platform (GeoMIP) was delivered to our government partners at the PANYNJ, by our industry partner VideoMining (formerly Advanced Interfaces). Working with PANYNJ has helped identify long-term tasks to make adoption of the GeoMIP system easier for crisis management. Examples include: (a) integrating the GeoMIP system with existing EOC software tools, (b) dynamically synchronizing GeoMIP system GIS data with master GIS datasets, and (c) providing role-specific, expertise collaboration to enable seamless group decision-making.

3.2 Supporting homeland security

The Pacific Northwest National Laboratory, through their Department of Homeland Security's National Visualization and Analytics Center, (<http://nvac.pnl.gov/>) or NVAC™, selected Penn State as the site for one of five new Regional Visualization and Analytics Centers (RVAC). The Penn State RVAC research builds directly upon our current GCCM project, a previous Digital Government project focused on development of visual analysis and communication methods, and the common NVAC goals to develop, implement, test, and deploy new visual analytics methods and technologies supporting the DHS mission.

The fundamental scientific objective underlying the PSU RVAC efforts will be understanding how individuals and teams carry out analytical reasoning and decision making tasks with complex information and using this understanding to develop and assess information technologies that enable these processes.

4. FUTURE CHALLENGES

In the coming year, will continue to implement and assess our approach to web-based geocollaboration for multi-organization response to crisis events, with a focus on international humanitarian logistics. We plan to leverage visual analytics methods that are developed through the RVAC introduced above, to extend the original capabilities planned for a collaborative, map-based web portal and to study use of this web portal to enable team work through extensions to our NeoCITIES simulation environment.

5. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grants No. BCS-0113030, EIA-0306845. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding agency.

6. REFERENCES

- [1] A. M. MacEachren, S. Fuhrmann, M. McNeese, G. Cai, and R. Sharma, "Project Highlight: GeoCollaborative Crisis Management," presented at 6th Annual National Conference on Digital Government Research: Emerging Trends, Atlanta, GA, 2005.
- [2] M. D. McNeese, "Pursuing medical human factors in an information age: the promise of a living lab approach," presented at Proceedings of Human Factors in Medicine, 2002.
- [3] I. Terrell, M. D. McNeese, H. Huang, S. Fuhrmann, and A. MacEachren, "The Use of Mobile Devices by West Nile Virus Field Workers," presented at Human Computer Interaction International, Las Vegas, NV, 2005.
- [4] R. E. T. Jones, M. D. McNeese, E. S. Connors, J. Tyrone Jefferson, and D. Hall, "A Distributed Cognition Simulation involving Homeland Security and Defense: The Development of NeoCITIES," presented at Human Factors and Ergonomics Society's 48th Annual Meeting, New Orleans, Louisiana, 2004.
- [5] R. M. Turner, "A Model of Explicit Context Representation and Use for Intelligent Agents," presented at Proceedings of the Second International and Interdisciplinary Conference on Modeling and Using Context, Lecture Notes In Computer Science, 1999.
- [6] G. Cai and Y. Xue, "Activity-oriented Context-aware Adaptation Assisting Mobile Geo-spatial Activities," presented at IUI'06, Sydney, Australia, 2006.

Indexing and Searching Handwritten Medical Forms

Venu Govindaraju

Center for Unified Biometrics and
Sensors, Department of Computer
Science and Engineering
University at Buffalo, NY 14260

govind@buffalo.edu

ABSTRACT

Extracting and reading handwritten data from medical forms is an important task in medical informatics as it paves the way for efficient archival, indexing, and retrieval. This paper addresses two important challenges: (i) extraction of handwritten text data from images of carbon copies, and (ii) intelligent use of context to reduce lexicons to make the task of handwriting recognition tractable. We have developed a smart binarization algorithm targeted to carbon copy images that outperforms methods reported in the literature. The lexicon reduction method is based on learning the medical concept, and hence the probable medical terms to be encountered in the narrative part that describes the chief complaint of the patient by training on OCR output. In our experiments, we have worked with about 600 medical forms, 20 medical concepts, and a lexicon size of 4,700. We have observed that if the concept is one of top 3 choices, the lexicon can be reduced by two-thirds on an unseen form.

Categories and Subject Descriptors

I.5 [Pattern Recognition]: Models –neural nets; Design Methodology – classifier design and evaluation, feature evaluation, pattern analysis; Applications- text processing. I.4 [Image processing and Computer Vision]: Segmentation

General Terms

Algorithms, Experimentation

Keywords

Medical Forms Processing, OCR.

INTRODUCTION

We will describe the progress on two tasks: (i) Binarization and text extraction from carbon copy form images, and (ii) lexicon reduction using medical concepts learned automatically.

TEXT EXTRACTION

We have evaluated several algorithms which extract handwriting from medical form images (Figure 1) to eventually provide the best handwriting recognition performance. The research copy of the NYS PCR [1] is a yellow-gray carbon mesh where both the handwriting and the mesh around the handwriting have approximately the same intensity. The absence of sufficient pen pressure causes character strokes to break after binarization which leads to recognition failures. Further, the broken/unnatural handwriting due to ambulance movement and emergency environments, and carbon smearing from unintentional pressure to the form add to the complexity of the binarization task. A lexicon driven word recognizer (LDWR) [2] is used for evaluation of the binarization methods.

Figure 1: Sample of a portion of the PCR with the form layout regions highlighted.

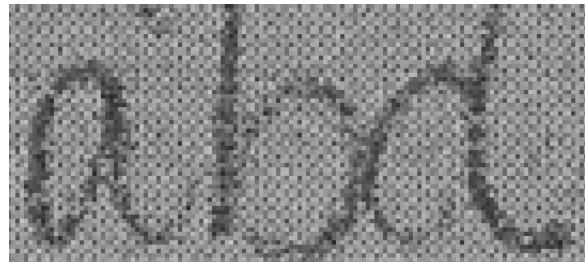


Figure 2: Zoom of text snippet highlighting the carbon mesh structure.

Figure 2 shows an example of the “Objective Assessment” region of the NYS PCR medical form. A zoom of ”abd” (a common abbreviation for abdomen) is shown to illustrate how the mesh and character strokes get fused. Prior algorithms have relied on techniques such as histogram analysis, edge detection, and local measurements. Our algorithm uses a wave trajectory for the positioned masks, as opposed to a linear trajectory (Figure 3).

A wave trajectory offers the benefit of a continuous trajectory of the masks regardless of distance from the starting point [3]. It allows the control of frequency and amplitude which are necessary to adjust for stroke width. Sinusoidal waves have been used in other contexts for the modeling of human motor function for on-line handwriting recognition, feature extraction, and segmentation. Intuitively, more space can be searched and both sides of the stroke can be evaluated in the same computational step at variable distances. It is also presumed that in a moving ambulance, carbon smearing is more likely since the writer will press harder to maintain balance in the vehicle.

Experiments were performed on 30 medical form images comprising of 1,440 word images and various size lexicons using the LDWR handwriting word recognition engine [2]. Our binarization algorithm followed by LDWR has 9-21% improvement, with various lexicon sizes, over Otsu binarization

method [4] followed by the same LDWR. Stop words are omitted. Additional 4-7% improvement is obtained by additional noise removal operations.

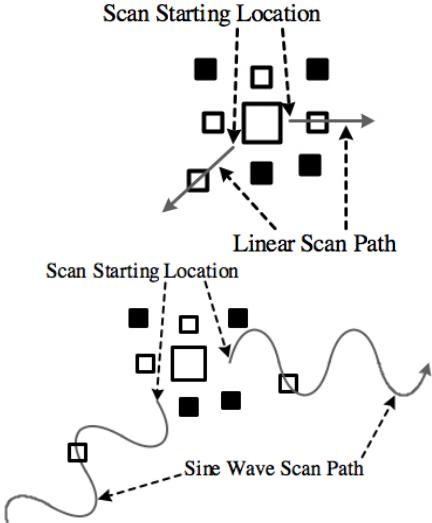


Figure 3: Scanning Approaches: Linear and Sinusoidal

LEXICON REDUCTION

A sequence of known characters (recognized with high confidence by an OCR), extracted from a form of initially unknown handwritten medical text, is used to represent a concept. Each concept contains a lexicon of words that have been encountered based on the training data. The concept is chosen as a human anatomical location where the patient ailment(s) exist. Multiple patient locations can be affected, however no PCR encountered has more than 5 locations and on average ~2 ailment locations were found on each form.

Example 1: A patient treated for an emergency pregnancy would be considered as the Reproductive System concept. Patients sometimes report additional pain to the Posterior Lumbar, Abdominal and Pelvic areas, which are additional concepts which could be assigned to the form.

Example 2: A conscious and breathing patient treated for gun shot wounds to the abdominal region would be considered Circulatory/Cardiovascular System, due to potential loss of blood, as well as other concepts such as Abdominal, Back, and/or Pelvic.

Each form is classified, by a human, as belonging to the highest priority anatomical concept. Words can belong to multiple concepts. The task is to automatically determine the relevant concept(s) which describe the content of a medical form. The reduced lexicon is defined as the summation of words from the determined concepts.

We adopt an approach similar to that used by Lucent Technologies Bell Laboratories to manage the call-routing problem [5]. Their navigation task was to direct a customer's call to a destination, in the domain of service offered, for a specific type of organization (e.g. banks). An automated system would construct a query from relevant n-gram terms in the user's dialogue. These n-grams were defined as a unigram, bigram and

tri-gram combinations of individual and adjacent relevant words from the human dialogue. The n-grams were modeled against correct destinations in the training stage and later used in the query to determine the target destination.

Given an unknown PCR form, the task is to determine the concept of the form, and use the reduced lexicon associated with the determined concept to drive the LDWR. Given a new PCR image, all image words are extracted from the form, and a word recognizer WR [6] is used to produce a list of confident characters for each word. For each word, only the most confident character and a pair of the most confident characters are considered. Next, nm-gram sequence lists [5] of adjacent words are composed using the confident characters from the WR.

PCR Training Document Size	550
PCR Testing Document Size	40
PCR Training Word Count	30,985
PCR Testing Word Count	1,603
PCR Training Term Count	51,803
PCR Testing Term Count	1,877/6,214
Anatomical Concepts Modeled	20
Average Words Per Document	55.23
Average Concepts Per Document	1.82
Complete Lexicon Size	4,700

If the top 1/3 of classified concepts are selected, 64% of the training set can benefit by eliminating 2/3 of the lexicon. Similarly, if the top half of classified concepts are selected, 70% of the training set can benefit by eliminating 50% of the lexicon

REFERENCES

- [1] V. Govindaraju and R. Milewski, "Automated reading and mining of pre-hospital care reports," in *IEEE Symposium on Computer-Based Medical Systems*, pp. 152-157, 2001.
- [2] G. Kim and V. Govindaraju: A Lexicon Driven Approach to Handwritten Word Recognition for Real-Time Applications. *IEEE Trans. Pattern Anal. Mach. Intell.* 19(4): 366-379 (1997).
- [3] R. Milewski and V. Govindaraju, "Extraction of handwritten Text from Carbon Copy medical Form Images", *Document Analysis Systems*, Nelson, NZ, 2006.
- [4] N. Otsu. A Threshold Selection Method from Gray-Level Histogram. *IEEE Transactions on System Man Cybernetics*, Vol. SMC-9, No. 1. C1979.
- [5] J. Chu-Carroll and B. Carpenter, "Vector-Based Natural Language Call Routing." *Computational Linguistics*. Vol. 25, No. 3, pp. 361--388, 1999.
- [6] [9] J. T. Favata: Offline General Handwritten Word Recognition Using an Approximate BEAM Matching Algorithm. *IEEE Trans. Pattern Anal. Mach. Intell.* 23 (9): 1009-1021 (2001).

Digital Governance and Hotspot Geoinformatics for Monitoring, Etiology, Early Warning, and Management Around the World

G. P. Patil

Penn State University

Department of Statistics

University Park, PA 16802 USA

Tel.: 001 814 865 9442

Email: gpp@stat.psu.edu

ABSTRACT

The five year NSF DGP project has been instrumental to conceptualize hotspot geoinformatics partnership among several interested cross-disciplinary scientists in academia, agencies, and private sector around the world. A declared need is around for statistical geoinformatics and software infrastructure for spatial and spatiotemporal hotspot detection and prioritization. Our efforts are driven by a wide variety of case studies of potential interest to government agencies involving critical society issues, such as public health, ecosystem health, sensor networks, robotic networks, social networks, video mining, homeland security, early warning, and disaster management.

Categories and Subject Descriptors

H. Information systems; H4. Information systems applications; H4.2 types of systems.

General Terms

Algorithms, Management, Measurement, Design, Security, Theory, Verification.

Keywords

Early warning, hotspot detection, hotspot geoinformatics partnership, hotspot prioritization, space-time hotspots.

1. INTRODUCTION

The primary thrust of the proposed work is now three-fold:

- a) To formulate and develop statistical methodology and computational technology for geoinformatic surveillance of hotspot detection and prioritization using upper level set detection and partially ordered set prioritization methods, software tools, and visualization capabilities.
- b) To formulate and initiate individual case study/application area project proposals that will have stronger and speedier performance, utilizing the detection and prioritization methods

and software tools of (a) above.

- c) To work toward a National Center for Geoinformatic Surveillance, utilizing (a) and (b) above as a synergistic springboard.

2. RESEARCH

Our research activities have focused on planning and software development for the methodological components of the geoinformatic surveillance project. The system has two methodological components: prioritization and hotspot detection.

3. SHORTCOURSES AND WORKSHOPS AROUND THE WORLD

You are invited to participate in the geoinformatics forum scheduled in various places and at various times around the world. The forum emphasis is on geoinformatics of hotspot detection and prioritization in a wide variety of subject areas and critical issues confronting agencies, academia, and industry. You are invited to participate in a manner most productive for your purposes, whether presentation of a paper with live case studies, attendance in a timely short course, or both. You will have the benefit of a veteran crossdisciplinary scientist as course instructor, workshop leader, and editor of resulting publications. And, of course, an opportunity to strengthen, advance, and accelerate your in-house research and work plan involving geoinformatics and hotspot dynamics. For information on Short Courses and Workshops, please see http://www.stat.psu.edu/hotspots/pdfs/OverallInfo_ShortCourseandWorkshops.pdf.

Two Days Short Course and Case Studies Workshop:

1. Siena, Italy (October 13-14, 2005)
2. Hyderabad, India (December 24-25, 2005)
3. Belo Horizonte, Brazil (April 6-7, 2006)
4. Nairobi, Kenya (March 18-19, 2006)
5. San Diego, USA (May 21-24, 2006)

One Day Short Course and Case Studies Workshop:

6. Okayama, Japan (November 19, 2005)
7. Bangkok, Thailand (November 30, 2005)
<http://www.j-geoinfo.net/HealthGIS/Symposium.htm>
8. Kuala Lumpur, Malaysia (December 27, 2005)
<http://iscm.math.um.edu.my/>
9. Jakarta, Indonesia (November 25, 2005 and January 9, 2006)

Course Instructor and Workshop Leader:

G. P. Patil

Distinguished Professor and Director,
Penn State Center for Statistical Ecology and Environmental
Statistics
Principal Investigator,
NSF Digital Government Research Project for Hotspot
GeoInformatics
Former Visiting Professor, Harvard School of Public Health
Editor-in-Chief, *Environmental and Ecological Statistics*
Fellow ASA, IMS, AAAS, RSS, ISI, IISA, NIE, DSEA
Administrative Information and Registration:
Nominal registration fees, if planned, will be reduced/waived
further for graduate research students, interested government
scientists and acceptable case studies investigators.

Contacts:

1. Lorenzo Fattorini

Universita degli Studi di Siena
Dipartimento di Metodi Quantitativi
P.zza S. Francesco 8
Siena 53100 ITALY
Telephone: 39/577-298624
Fax: 39/577-298626
Email: fattorini@unisi.it

2. C. R. Rao

C. R. Rao Advanced Institute
for Mathematics, Statistics, and Computer Science
Osmania University
Hyderabad, 500 007, INDIA
Department of Statistics
Penn State University
Email: crr1@psu.edu

3. Renato Assuncao

Universidade Federal de Minas Gerais
Departamento de Estatistica
Instituto de Ciencias Exatas
Campus Pampulha
Belo Horizonte MG 31270 901 BRAZIL
Email: assuncao@est.ufmg.br

4. Ashbindu Singh

Regional Coordinator
UNEP Division of Early Warning and Assessment
Washington, DC 20006 USA
Telephone: 202/785-0465
Fax: 202/785-2096
Email: as@rona.unep.org

5. G. P. Patil

Email: gpp@stat.psu.edu

6. Koji Kurihara

Okayama University
Faculty of Environ. Sci. & Technology
2-1-1 Tsushima-naka
Okayama 700-8530 JAPAN
Telephone: 81 86 251-8508
Fax: 81 86-251-8552
Email: kurihara@ems.okayama-u.ac.jp

7. Nitin Tripathi

Editor-in-Chief, International Journal of Geoinformatics
Chairman, Association of Geo-Information Technology (AgIT)
School of Advanced Technologies
Asian Institute of Technology
P.O. Box: 4,
Klong Luang, Pathumthani 12120 Thailand
Telephone: 66-2-5246392
Fax: 66-2-5245597
Email: nitinkt@ait.ac.th

8. S. H. Ong

Institute of Mathematical Sciences
University of Malaysia
50603 Kuala Lumpur, Malaysia
Email: Ongsh@um.edu.my

9. Asep Saefuddin

Vice Rector of Bogar Agricultural University
for Planning, Development, and Collaboration
Kampus IPB Darmaga Bogor, 16680
Indonesia
Telephone: 62 251 622 643
Fax: 62 251 624 057
Email: wakilrektor4@ipb.ac.id

4. ACKNOWLEDGMENTS

This material is based upon work supported by (i) the National Science Foundation under Grant No. 0307010, (ii) the United States Environmental Protection Agency under Grant No. CR-83059301 and (iii) the Pennsylvania Department of Health using Tobacco Settlement Funds under Grant No. ME 01324. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the agencies.

5. REFERENCES

- [1] Patil, G.P., Geoinformatic Hotspot Systems (GHS) for Detection, Prioritization, and Early Warning. Project highlights, dg.o2005, Atlanta, Georgia.
- [2] Patil, G.P., and Taillie, C. Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statistics*, 11 (2004), 183-197.
- [3] Patil, G.P., and Taillie, C. Multiple indicators, partially ordered sets, and linear extensions: Multi-criterion ranking and prioritization. *Environmental and Ecological Statistics* 11 (2004), 199-228.
- [4] Patil, et al. Upper level set scan statistic system for detecting arbitrarily shaped hotspots for digital governance. Poster presentation, dg.o2005, Atlanta, Georgia.
- [5] Patil, et al. Geoinformatic surveillance of hotspot detection, prioritization and early warning. Demo, dg.o2005, Atlanta, Georgia
- [6] Patil, et al. Hotspot geoinformatics for digital governance. Encyclopedia of Digital Government, Ari-Veikko Anttiroiko and Matti Malkia (eds.), 2006, to appear.

SESSION 3B

CITIZEN PARTICIPATION 2

Moderator

Thomas Horan, Claremont Graduate Institute, USA

Titles and Authors

When Opinion Leaders Blog: New forms of citizen interaction
Kavanaugh, Andrea; Zin, Than Than; Carroll, John M.; Schmitz, Joseph; Pérez-Quiñones, Manuel; Isenhour, Philip

Bringing an Informed Public into Policy Debates through Online Deliberation: The Healthcare Dialogue project
Price, Vincent; Cappella, Joseph N.

Decoding Political Discourse Networks
Kelly, John; Stark, David

When Opinion Leaders Blog: New forms of citizen interaction

Andrea Kavanaugh +
kavan@vt.edu

Joseph Schmitz †
J-Schmitz@wiu.edu

Than Than Zin +
tzin@vt.edu

Manuel Pérez-Quiñones +
perez@vt.edu

John M. Carroll ±
jcarroll@ist.psu.edu

Philip Isenhour +
Isenhour@vt.edu

+ Department of Computer Science, Virginia Tech, Blacksburg, VA 24061-0106; (540) 231-1806

± School of Information Sciences & Technology, Penn State, University Park, PA 16802; (814) 863-2476

† Department of Communication, Western Illinois University, Macomb, IL 61455; (309) 298-2370

ABSTRACT

Web logs (i.e., blogs) provide enhanced opportunities to extend capabilities of traditional electronic mail and discussion lists, especially in the hands of opinion leaders; such tools offer greater social interaction and informal discussion, and opportunities for conversational content production. Because blogging tools are simple, available, and free, users can easily communicate with others in their social networks, their geographic communities and the interested public. Blogs represent self-organizing social systems that can help many persons to: 1) interact collaboratively, 2) learn from each other by exchanging ideas and information, and 3) solve collective problems. For opinion leaders – that small percentage of the population that is socially and politically active – blogs represent another channel to disseminate ideas and garner feedback from members of their social network. The present research offers findings from a random household survey of citizens of Blacksburg and Montgomery County, Virginia about citizens' interests and attitudes towards local government, discussion of political issues, and their Internet use. We find that opinion leaders who engage in some form of blogging (read or write) are more likely to be male, extroverted and educated than bloggers who are not politically active. They score higher than other bloggers on measures of offline and online political interests and activities, community collective efficacy, and the size and heterogeneity of their political discussion networks. As such, their use of blogs may serve as a growing new communication channel to exercise their informal influence.

Categories and Subject Descriptors

H.5.3 [Information Interfaces And Presentation]: Group and Organization interfaces - *collaborative computing*.

General Terms

Measurement, Human Factors, Theory.

Keywords

Social computing, Internet, computer mediated communication, empirical methods, survey research.

1. INTRODUCTION

For at least three decades, scholars and information technology experts have envisioned computer-based communication technologies as a basis for a vibrant, engaged, and informed democracy (Arterton, 1987; Horrigan, 2001; Rogers, 1986; Schuler, 1996). Although many of these utopian hopes have not been realized, computer networking has fostered greater participation in democratic political life in the United States (Barber, 1984; Coleman and Gotz, 2002; Kavanaugh et al., 2005a, 2005b; Rainie, 2005; Schmitz, et al., 1995).

Electronic mailing lists and politically oriented web-based resources grew rapidly in the late 1990's. Much of this information and communication technology (ICT) enhanced participation has: 1) increased awareness about issues and problems, and 2) increased capabilities for coordination, communication and outreach for political actors. Recent large-scale examples include the 1999 coordination of demonstrations against the World Trade Organization meeting in Seattle that were greatly facilitated by online communication. During the 2004 US presidential campaign, political groups coordinated such activities as leafleting, neighborhood canvassing, and fundraising over the Internet. Thus, when campaign volunteers from Southwest Virginia sought to help distribute surplus campaign materials in Ohio, they went online to find their counterparts, establish contact with them, and coordinate rendezvous points.

These technologies (ICT) have a long history of utilization in local political activities dating from early campaigns to change how a city would treat its homeless waged on Santa Monica's Public Electronic Network (Schmitz et al., 1995) or the establishment of neighborhood coalitions via the Seattle Community Network (Schuler, 1996). When a group of citizens in Blacksburg, Virginia documented controversial procedures surrounding a proposed local sewer development, they used email and listservs, along with face-to-face campaigning to ensure a record turnout of citizens that elected sympathetic candidates as new Town Council members in a landslide victory (Kavanaugh et al., 2005b).

In short, the enhanced capabilities of ICT to reach and mobilize more people, more quickly and to garner greater resources has

helped political actors to raise awareness, educate citizens, mobilize supporters, and coordinate collective responses. Much of the ICT use for political awareness and mobilization has been dominated by what are commonly known as ‘opinion leaders’ or ‘influentials.’ Opinion leaders are characterized by higher social status (even within lower social strata) and gregariousness (talkative, extroverted, and affiliated with more clubs and local groups). They are also more exposed to and responsive to new information and ideas, traditionally through various mass communication media (Katz and Lazarsfeld, 1955; Keller and Berry, 2003, among others). Katz and Lazarsfeld’s (1955) ‘two-step’ flow of communication model has been well established through numerous subsequent studies: that is, ideas often flow from radio and print to opinion leaders and from them to the less active segments of the population.

The use of new communication technologies, such as blogs, by opinion leaders raises such questions as: To what extent do opinion leaders use blogging for political discussion, opinion sharing, and information exchange? Is there evidence that political opinion leaders (who by definition have large discussion networks, i.e., many friends, family and acquaintances with whom they discuss politics) are likely to use blogs to supplement face-to-face discussions? There is the potentiality at least that as a result of blogging by opinion leaders, individuals oriented to these influentials may become drawn into a more deliberative type of participation based on increased opportunities for informal discussion and exchange of ideas.

2. PRIOR RESEARCH

We have been investigating citizen participation in governance at the local level in order to understand the extent to which existing information and communication technology (ICT) satisfies community communication and information needs and interests during a program of research that has extended for more than five years and has included many researchers with multi-disciplinary interests. At the local level we have asked (and answered) such questions as: Who is politically active? What is the nature of their political involvement? How do they use ICT?

The overall goal of our research program—and much other the research about information and communication technology and political participation—is to study the effects of ICT use upon the level and type of participation by different citizens. We are especially interested in the role of opinion leaders and their use of new information and communication technology.

2.1 Opinion Leadership

Opinion leadership, summarized by Rogers and Shoemaker (1971) is “the degree to which an individual is able to informally influence other individuals’ attitudes or overt behavior in a desired way with relative frequency” (p. 35). The influence is informal, often by word of mouth, at the local level with ideas and information spreading throughout a leader’s social circle (Keller and Berry, 2003). Further, opinion leadership is earned and maintained by the individual’s competence, social accessibility, and conformity to agreed upon norms. Opinion leadership theory argues that opinion leaders (or ‘influentials’ as they are also called) exist at all social strata and can vary somewhat by subject area (politics, technology, consumerism) although they are generally attuned to new ideas and forward thinking across the board. Extensive longitudinal research by Roper (since the early

1970’s) shows that influential Americans have consistently been politically aware and socially active citizens (Keller and Berry, 2003).

In 30 years of survey research, Roper has found that to qualify as an influential, a person had to have done three or more of a list of social or political activities in the past year (as self-reported in surveys), such as: written or called any politician at the state, local or national level; attended a public meeting on town or school affairs; held or run for political office; served on a committee for some local organization; served as an officer for some club or organization; written a letter to the editor of a newspaper or magazine (or called a live radio or TV show to express an opinion); signed a petition; worked for a political party; made a speech; written an article for a magazine or newspaper; or been an active member of any group that tries to influence public policy or government (Keller and Berry, 2003, p. 19).

We used this Roper set of questions in identifying respondents as opinion leaders in our survey findings reported here, building on our prior research on “leader and member bridges” discussed further below. Ideally, we would have corroborated that other community members also identified these same politically and socially active survey respondents as opinion leaders. However, this would have required more extensive interviews with local residents than we were able to accommodate. The lack of such corroboration is a limitation of our study. Nonetheless, Roper consistently found that a person who had done three or more of the social or political activities listed (above) in the past year was also identified by others in interviews as an influential. We add to Roper’s minimum survey measures for social and political activism a measure of extroversion (i.e., sociability and gregariousness) plus a measure of the number of local groups with which the respondents were affiliated. Therefore, taken altogether, our measures for identifying influentials in our sample seem reasonable and adequate.

Opinion leaders make up only about 10-15 percent of the total US population and they are that same group that is comprised of the politically active citizens (Verba and Nie, 1972; Milbrath and Goel, 1977; Dahl, 1991; Norris, 2001; among others). Thirty years of Roper research shows that influentials were among the early adopters of information technology, including video recorders and home computers. Their early use of applications like word-processing led them see the benefits of personal computing ahead of the general population (Keller and Berry, 2003). Influentials have been using the Internet, including email, bulletin board systems, and web browsers from the outset to stay informed and involved in political issues that interest them at local, national and international levels. Since people turn to them for advice and opinions, influentials’ use of online information resources helped them spread outside ideas throughout the general population in the course of face-to-face political discussions and computer mediated communication opportunities (e.g., email and listservs).

The use of traditional ICT (such as, email and listservs) by political activists (most of which are opinion leaders) to help inform, involve, and mobilize citizens for collective action are well documented (Dahlberg, 2001; Norris, 2001; Horrigan, et al., 2004, among many others).ⁱ National survey research and comprehensive case studies have also demonstrated the

effectiveness of ICT to increase civic awareness and participation among interested citizens.

The ‘word of mouth’ and conversational style of blogs makes them particularly well suited to the often informal style of influence used by opinion leaders. The design and usability of blogs harken back to some of the earliest forms of Internet-based social applications, such as bulletin board systems, where interaction and exchange of ideas were more prominent online than information browsing. The role that influentials play in disseminating and discussing ideas most often informally and by simple ‘word of mouth’ (and, more recently, augmented by email and listserv) is essential to deliberative democracy.

2.2 Deliberative Democracy

Deliberative democracy is a political system based on the open public discussion and consideration of political ideas and problems with a view to collective opinion formation, decision-making and response (Barber, 1984; Fishkin, 1991, among others). Underlying institutional procedures, such as rule of majority, is a culture of political discussion and voluntary participation. In the view of Kim, Wyatt and Katz (1999) citizens’ free discussion of public issues is the soul of democracy. Their interest is not so much with formal discussions with specific agendas and purposes, but rather the casual, off-hand and spontaneous conversations that spring up routinely throughout the average person’s daily life. Their conversation may generally occur in the private sphere (with family members and friends), but its substance (e.g., information and ideas) comes from outside (e.g., media and opinion leaders). While not everyone agrees that casual political conversation is valuable to democracy (see Schudson, 1992, 1997, for example), who argues that goal-oriented discussion that is guided by rules is more fruitful and valuable for democratic talk.

Several innovative projects have rule-based formal public forums online for citizens to discuss important issues and to deliberate on a variety of national, global and local policies (e.g., Minnesota E-Democracy, UK E-Democracy, and many other specialized discussion groups). These online deliberation forums provide valuable central sites for public participation in policy debate and decision-making (Gastil and Levine, 2005; Hill and Hughes, 1998; Katz and Rice, 2002; Kirn, 2002; Price, 2005). Deliberative polling offers effective experiments to understand the impacts of information and deliberation upon citizens’ opinion formation and consensus building (Fishkin, 1991).

A critical limitation of centralized sites/forums to discuss issues is that they tend to attract the “usual” activists who are presently comfortable with computer technology. Existential centralized sites are difficult to scale up and thus reach persons beyond a core group of activists. Presently, forum leaders expend time and effort to attract and recruit participants to their sites since most potential participants are not highly motivated to seek these special locations where they might air their views and concerns. Yet, citizens who may be ‘passive supporters’ are not without opinions and they may well express their political views to their network of friends, family and acquaintances (both offline and online). Often ‘passive supporters’ express their opinions online via email with friends and family or they may use mailing lists to which they are subscribed.

Early adopters are beginning to use some of the newer technologies, such as blogs and wikis (perhaps casually, yet

importantly) to express their personal and political views. As we try to show here from our research findings, among these early adopters are opinion leaders. Blogs are easy to set up and use. They are free and widely available through several services such as livejournal.com, xanga.com, and blogspot.com. Low barriers to entry and easy content authoring have spurred their recent rapid growth. National statistics on blogging (Rainie, 2005) indicate that by the end of 2004, about 27% of Internet users reported reading blogs, while 7% of Internet users had reported creating a blog.

Motivations for blogging range from the intimately personal to the globally political (Nardi et al., 2004; Rainie, 2005). Nardi’s ethnographic study of “ordinary” bloggers, identified five major motivations: 1) documenting one’s life, 2) providing commentary and opinions, 3) expressing deeply felt emotions, 4) articulating ideas through writing, and 5) forming and maintaining community forums. She also found that the motivations for blogging may overlap. For example, one blog was for a class whose professor noted: “We’ll try to take advantage of the general nature of Weblogs as ‘public journals’ in using them for personal reflection in the context of a learning community, on issues that arise in the course, both rhetorical and content-related.” (Nardi, 2004, p. 45) In Nardi’s view, this professor hoped to “facilitate the building of the learning community by getting students in conversation with each other electronically.” (p. 45). Community forums for geographic locales such as Burlingame, California (<http://www.burlingame.org>) seek similar self-organizing social systems (Wiley and Edwards, 2003) in which people discuss local problems with each other and help to solve them, collectively.

2.3 Blogs as Self-Organizing Social Systems

Blogging software distributes responsibility for content creation, commentary, and quality control across a community of users (i.e., writers, commentators, and readers). As such, blogs offer potential frameworks for effective deliberation and thus, provide users with critical tools needed for self-organization. Self-organizational models for human behavior help us to understand urban planning (Jacobs, 1961), economics (Krugman, 1996), organizational structures (Wheatley, 1992), and computer supported cooperative work (Wulf, 1999). For example, Wulf investigated ways that ‘groupware’ systems are used to support self-organization.

When a critical mass (Markus, 1987) of politically active, Internet users adopt such innovative technologies as web logs for political discussion, they create online self-organizing systems for democratic purposes that may become self-sustaining. In this way “outsiders” are no longer required to recruit people to join a centralized formal on-line discussion. Blogging provides needed tools for people who are presently discussing civic life (among other things) that can help communities of users organize and shape discussions among themselves providing interest is maintained and subjects remain timely.

2.4 The Blacksburg Electronic Village

Blacksburg and Montgomery County in southwest Virginia offer a rich opportunity to investigate Internet use. Blacksburg hosts the land grant state university of Virginia Polytechnic Institute & State University (Virginia Tech) and is home for the mature, well-established community computer network known as the Blacksburg Electronic Village (BEV). Although its name implies otherwise, the BEV does serve the County in which Blacksburg is

located and extends services to a wider region beyond the local planning district. The population of Blacksburg (estimated 38,000 in 2005) is largely affiliated with Virginia Tech as faculty, staff, or students. Nearby Christiansburg, with a population of 22,000, houses a mixture of Virginia Tech affiliates and working class households.

Virginia Tech, in partnership with the Town of Blacksburg and the local telephone company (then Bell Atlantic, now Verizon), launched the BEV in 1993. The university provided Internet access through its modem pool to residents and supported a small staff whose goal was to develop web-based local content and build a critical mass of users through user training and support. By 1995, random sample surveys of Blacksburg households indicated that 62% of the respondents were using the Internet (Kavanaugh, et al., 2000). In addition to residential users, the town government, county government, county public schools, public health offices, and county public libraries also maintained content-rich web sites (hosted on BEV servers); these agents also intensively trained their personnel to use this unique community network. The high levels of institutional and residential users provided an attractive market for private Internet Service Providers who then offered dial-up and high speed connectivity (DSL, Ethernet in apartments, and eventually cable modem) to area residents.

Random sample household surveys since 1995 have shown a steady rise in the number of Blacksburg respondents reporting they use the Internet through 2001, when at 89%, penetration reached a saturation point. That is, everyone who wanted to be online was, and those who were not online had either chosen not to use the Internet or were using email and web resources through surrogates (often family members or friends). Unlike Blacksburg residents, the number of people in surrounding Montgomery County who reported they used the Internet use has steadily risen (from about 20% in 1999 to 68% in 2005). For user populations in both locations (Blacksburg and Montgomery County) the *amount* of daily usage and *types* of usage have changed over time, particularly as more choices in broadband technologies (e.g., cable modem, satellite connectivity) and new developments in software programs and applications have been offered. Most local residents and organizations use the Internet routinely; most agents also *expect* that almost everyone else can access information and communicate with each other online.

3. RESEARCH METHODOLOGY

Through in depth interviews, archival records and web searches, we have been investigating current information and communication technology use and practices among local citizens, community groups and government representatives. We also investigated ICT use among other communities in the United States, particularly where we identified active online use of community discussion forums, blogs and/or wikis (Kavanaugh, et al., 2005c). We designed and administered a random sample household survey to assess political participation and Internet use in Blacksburg and Montgomery County, Virginia. We were particularly interested in those respondents who had heard of blogs, and were reading, posting comments to, or writing blogs. We report here highlights of the findings of the present research.

3.1 Household Survey

We developed the survey instrument by selecting questions from our prior BEV research, the HomeNet study (Kraut, et al., 1996, 2002) and validated questions regarding political efficacy, participation and attitudes that have been used in many prior studies (e.g., Dahl, 1991; Michaelson, 2000; Miller, et al., 1980; Verba and Nie, 1972). We contracted with Virginia Tech's Center for Survey Research to 1) transform the questionnaire into a telephone interview format, and 2) to conduct interviews by telephone with a sample of 1200 households. Our sample consisted of households in Montgomery County, including the two large towns of Blacksburg and Christiansburg that lie within the county limits. We had purchased the sample from Survey Sample, Incorporated who generated our random sample from listed and unlisted telephone numbers available to Montgomery County, Virginia residents. After eliminating all ineligible records (e.g., outside Montgomery County, hearing disabilities, etc.), the number of eligible sample members was 1,795. A total of 717 interviews (response rate of 40%) were completed.

3.2 Survey questions and constructs

All constructs for this project were represented by variables that were subjected to reliability analysis. The present survey questions were selected so that the constructs were as similar as possible to those in our prior study about community computing in Blacksburg and Montgomery County: Experiences of People, Internet and Community (EPIC) reported elsewhere (Carroll and Reese, 2003; Kavanaugh, et al., 2003; Kavanaugh, et al., 2005a; Carroll, et al., 2005). Primary constructs included the following:

Offline Political and Civic Interests How frequently in the last six months the respondent read local, national and global news in the paper; attended a local public or political talk or meeting; wrote or called a local government official; attended a religious service; did volunteer work;

Online Political and Civic Interests How frequently in last six months the respondent used the Internet for the following: to work for a political party or candidate, to try to influence a government policy or affect a politician's point of view; to send email to a local government official; to get local, national or global news; to read, comment on or write a blog; to post factual information for citizens; to express opinions in forums or group discussions;

Internet Helpful for Involvement Level of agreement with 1) The Internet has helped me feel more connected with people like myself in the local area; 2) The Internet has helped me feel more connected with a diversity of people in the local area; and 3) The Internet has helped me become more involved in local issues that interest me.

Political efficacy Level of agreement with the following: 1) Sometimes local politics and government seem so complicated that persons like me can't truly understand what's going on, 2) I don't think local public officials care much what people like me think, and 3) There are plenty of ways for people like me to have a say in what our local government does.

Community Collective Efficacy Level of agreement with the statement "I am convinced that we can improve the quality of life in the local community, even when resources are limited."

Trust Composite variable comprised of the following: To what extent do you think most people in the local area can be trusted?

And To what extent do you think most people in the local area are inclined to help others?

Extroversion Level of agreement with the statements 1) "Generally speaking, I am outgoing and sociable; and 2) "I am talkative."

Political Talk Frequency measures on the following: In the last six months (how frequently) have you 1) talked to family members about local issues or concerns; 2) talked to family members about national or global issues; 3) talked to people outside your family about local issues; and 4) talked to people outside your family about national or global issues.

Our political efficacy construct is based on questions used in prior research (Michaelson, 2000; Miller, et al., 1980; Verba and Nie, 1972, and others), where political efficacy is defined as the belief that individual political action does have, or can have, an impact upon the political process. Since this definition does not include the notion of obstacles that must be overcome, as Bandura's (1997) concept of efficacy requires, it comes closer perhaps to a sense of political empowerment than efficacy in Bandura's sense. Our construct combines both internal and external political efficacy. Internal efficacy "indicates individuals' self-perceptions that they are capable of understanding politics and competent enough to participate in political acts such as voting" (Miller et al., 1980, p. 253). External efficacy "measures expressed beliefs about political institutions. The lack of external efficacy ... indicates the belief that the public cannot influence political outcomes because government leaders and institutions are unresponsive."

Our question we name 'community collective efficacy' is taken from a larger construct we developed extensively in the earlier EPIC research referred to as collective efficacy (see Carroll and Reese, 2003). We mean specifically a feeling of efficacy about the capacities of one's community to achieve goals in the face of difficulties or limitations.

Space does not permit us to provide full details on all these constructs; we have posted more information (name, label, range and average values, number of valid cases, component variables, their correspondent survey questions, and value labels for each measurement scale) at <http://java.cs.vt.edu/public/projects/digitalgov/data>.

We created a *Political Discussion Network* construct ($\alpha = .75$) that included the *Political Talk* construct described above and seven additional variables: 1) number of people outside family with whom respondents talk about an issue they consider to be the most important facing the local area; 2) likelihood of attending meetings on this issue, 3) likelihood of speaking at a public forum on this issue, 4) likelihood of expressing a different opinion at a public forum on this issue; and, in the last six months, how frequently the respondent has: 1) attended a local political talk or meeting, 2) attended a public meeting and 3) discussed politics. The *Political Discussion Network* construct is composed of variables' z-scores since the different questions used different scale metrics.

Primary questions referred to family, friends, or acquaintances with whom respondents discussed politics (local, national or global politics). We included a question in this construct respondents' likelihood of attending a public forum on a specific local issue that they identified as the most important issue facing

the community. We also asked about the likelihood of their expressing opinions about this issue during a public forum. Additional measures captured the size of respondents' political discussion networks by asking about how many people outside their immediate family had they discussed their most important local issue (the issue they had earlier identified).

This paper analyzes differences between bloggers and respondents who had never heard of blogs, and between politically active 'bloggers' and politically inactive 'bloggers'. We calculated the composite variable "blogger" as the sum of four variables: 1) heard of blogs, 2) read blogs, 3) posted comments to blogs, and 4) wrote blogs. The variable 'heard of blogs' had two response categories: yes=1, no=0. The remaining three questions used a frequency scale ranging from 1=never to 6= several times a day. We collapsed these six response categories into three: 0=never, 1= occasional, and 2= frequent. The sum of scores across these four questions, using recoded response categories yielded a scale that ranged from 0 to 7. The sub set of the sample (319 respondents) who had at least heard of blogs is a basis of the blogger analyses we describe below.

Neither the level of blogging experience (i.e., heard of blogs, read, post and/or write them), nor the frequency of blog activity (e.g., several times a day) discriminated between respondents who are interested in more political versus more personal types of content. Therefore, we subdivided 319 respondents who had at least heard of blogs (what we call the 'blogger' population) into those who had larger versus smaller political discussion networks. Using their scores on the Political Discussion Network construct, we discriminated between two groups: political bloggers (N=93) and personal bloggers (N=226) based on a cut-off point at the 75th percentile (using a score of 0.39 or above on the 'political discussion network' measure). For the remainder of this paper we refer to bloggers with larger political discussion networks simply as political bloggers, and those with smaller or no political discussion network as personal bloggers.

We tested bivariate correlations to explore differences between political and personal bloggers and we used *t* tests to examine differences more rigorously. We also used *t* tests with primary survey constructs and some demographic variables to identify significant differences between light versus heavy bloggers (i.e., based on frequency of blogging activity). The characteristics, interests, and activities of people who are politically active and who are also somewhat familiar with blogs offers unique insights about how this new information technology presently supports and/or enhances individuals' civic engagement.

4. RESULTS

Completed surveys were geographically representative compared to our original sample; that is their distribution matched what one would expect from Blacksburg, Christiansburg, and the remaining portions of Montgomery County. Almost eighty percent (78%) of the total respondents reported they use the Internet. Internet users tended to have higher levels of education, household income, and household size (more children living at home). Internet users were more likely to be affiliated with more formal and informal groups than non-Internet users (Table 1). In all the tables below, due to limitations of space, we do not report the non-significant findings.

Table 1. Internet Users: Demographics and Attributes

Variable Names	Mean (SD)	Valid N
Age	44.4 (15.37)	552
Location	1.86 (0.77)	556
Education	4 (1.4)	553
Household income	1.58 (0.49)	518
Children at home	0.43 (0.49)	495
Formal groups	1.15 (1.26)	554
Informal groups	0.70 (1.02)	553

Correlations were calculated between responses to the question "Do you use the Internet from any location" and selected demographic characteristics to determine if that Internet use was associated with gender, age, estimated household income, home ownership, household size, education, marital status, living with children, extroversion, number of formal and informal group affiliations, and a one-item measure of community collective efficacy. Table 2 presents statistically significant (Kendall-tau) correlation coefficients.

Table 2. Correlates of Internet Use Variables

Variable Name	Correlations	Valid N
Age	-.323**	709
Location ⁱⁱⁱ a	-.175**	716
Household income ^b	.298**	661
Education ^c	.418**	712
Marital status ^d	-.209**	713
Children at home ^e	.111**	602
Household size	.202**	715
Formal group affiliations	.166**	713
Informal group affiliations	.199**	703

* p<.05, ** p<.01

We examined relationships between the primary survey constructs and demographic variables (e.g., age, residential location, household income, home-ownership, number of people living at home, level of formal education, marital status, households with children under the age of 18, and extroversion). We also tested for relationships with variables that capture our respondents' identification of and reactions to important local issues (Table 3).

Table 3. 'Most Important Issue Facing Local Area'

Variables: Most Important Issue	Constructs	Correlation (N)
Important local issue provided (or not)	Offline political interest	0.21** (618)
	Offline civic interest	0.21** (618)
	Political talk	0.22** (618)
Likely to attend public forum on important issue	Offline political interest	0.23** (512)
	Political talk	0.24** (512)
	Offline political activities	0.31** (512)
Likely to speak at public forum on the important issue	Extroversion	0.26** (420)

* p<.05, ** p<.01

Table 3 shows significant correlations between the main survey constructs and whether the respondent provided what they considered to be the most important local issue, as well as how many persons outside of family members with whom s/he talked about the most important local issue, whether these people shared his/her point of view, the likelihood of their attending a public forum on the their issue, the likelihood of their speaking at that public forum, and their preferred type of news source. One effect of the relatively large sample size was that many of the relationships were statistically significant. Table 3 reports constructs that were moderately correlated ($r>0.20$, Pearson correlation coefficients) with the responses to questions that address "the most important issue" given by respondents.

Regarding blogs, among Internet users, 44.5% (319 people) reported they had heard of blogs. Among this subset of the population who had heard of blogs, more than half (57%) had never read blogs. Of the remaining 43% who read blogs, the largest proportion (23.3%) rarely read them (about once a month or less). Only about one-fifth (19.8%) reported reading blogs once a week or more.

An even smaller proportion (18.9%) of those people who had heard of blogs reported ever posting comments to them. Further, most persons who had posted comments on blogs reported posting infrequently (less than once a month). Only 7.2% of those who had heard of blogs had posted comments about once a week. Not surprisingly, authors of blogs were the most rare. Only 7% of Internet users in our study were blog writers (a finding similar to national statistics from Horrigan, et al., 2005, gathered at the end of 2004). Correlation findings suggest that respondents who have at least heard of blogs (but may also read, post and/or write blogs) use the Internet for different purposes than do those respondents who have not yet heard of blogs.

Specifically, the more experience people had with blogs, the more often they used the Internet to: 1) get national and global news, 2) look for information on the BEV web site, 3) work for a political party, 4) influence policy, 5) post information for other citizens, and 6) express their opinion in an online forum. Respondents with minimal blog use but who had at least heard of blogs used the Internet more frequently than Internet users who had not heard of blogs to: 1) get national and global news, 2) look for information on the BEV web site, 3) post information online for other citizens, 4) express their opinions in online forums, and 5) use the Internet to influence policy. See Table 4.

Table 4. 'Heard of blogs' and Frequency of Internet Use

Variable Name Frequency of Internet use to:	Pearson Correlations	Valid N
Get national/global news	0.30**	555
Get info from BEV website	0.08*	552
Work for political party	0.12**	555
Post information for citizens	0.13**	554
Express opinion in online forum	0.19**	555
Influence policy	0.15**	555

* p<.05, ** p<.01

Among Internet users, we scored each respondent according to their level of knowledge and experience with blogging. Scores ranged from zero, for someone who had never heard of blogs, to a high of seven for someone who writes a blog about once a week or more. A higher score also included a measure that captured

reading and/or commenting on blogs. Results of *t* tests comparing respondents with different levels of blogging experience showed that persons with more experience with blogs also scored higher on measures of Internet experience and with the amount of Internet use on a typical day (Table 5).

Table 5. Differences Among Respondents by Blogging Awareness/Experience

Variable Labels	Means		<i>t</i> test (df)
	Not Aware	Aware User	
Number of years using Internet	7.9	9.4 (548)	-4.50**
Hours/day spent on Internet	1.9	2.6 (539.25)	-3.87**

* $p < .05$, ** $p < .01$

Respondents who heard of/read/posted to and/or wrote blogs were more interested and active in politics and civic life, both online and offline, and they discussed politics with more people than respondents who had never heard of blogs. See Table 6.

Table 6. Differences Among Respondents by Blogging Awareness/Experience

Variable Labels	Means		<i>t</i> test (df)
	Not Aware	Aware User	
Offline political	1.33	1.45	-2.55** (548.89)
Online political	1.12	1.24	-3.65** (552.26)
Online civic	1.68	1.82	-2.73** (553)
Political talk	3.61	3.93	-2.98** (468.9)
Informal groups	0.59	0.78	-2.19* (549)

* $p < .05$, ** $p < .01$

One important finding was that people with more experience with blogs (readers, commenters, writers) were more likely to belong to a greater number of *informal* groups (the most common of which were social and recreational types of groups). When we subdivided people who had at least heard of blogs into those who were more politically involved and those who were less politically involved, we found that bloggers who were politically and socially active (i.e., opinion leaders) belonged to more *formal* groups. Politically active citizens who engage in some form of blogging were different from ‘personal’ bloggers (Tables 7-9 and Notes).ⁱⁱⁱ

Table 7: Bloggers: Demographics and Attributes

Variable Labels	Means		<i>t</i> test (df)
	Personal	Political	
Age	43.34	47.01 (191.44)	-1.98*
Extroversion	3.08	3.40 (223.12)	-3.73**
Estimated household income	1.57	1.71 (166.2)	-2.45*
Home ownership	0.70	0.82 (200.92)	-2.27*
Number of people living at home	2.52	2.86 (317)	-2.40*
Number of formal group affiliations	0.92	1.71 (137.36)	-4.99**

* $p < .05$ ** $p < .01$

There were significant differences on measures of demographics, attitudes, interests, activities, and political discussion networks. Politically active citizens who use blogs tended to be slightly older (average age 47), male, more extroverted, and better educated (average education=some graduate school) as shown in Table 7. As a category, therefore, we consider ‘politically active’ bloggers to be opinion leaders who engage in some form of blogging (even though it may just be reading them). Political bloggers also scored higher than personal bloggers on measures of offline and online political interests and activities, community collective efficacy, and on the size and heterogeneity of their political discussion networks. See Table 8 (the number of political bloggers ranged from 84 to 93).

Table 8: Political vs Personal Bloggers Attitudes/Interests

Variable Labels	Means		<i>t</i> test (df)
	Personal	Political	
Community collective efficacy	3.33	3.51 (312)	-2.21*
Offline political interest	1.27	1.89 (120.37)	-7.40**
Offline civic interest	3.14	3.66 (317)	-4.82**
Political talk	3.46	5.07 (233.04)	-16.22**
Offline political activities	1.27	1.77 (122.50)	-8.703**
Online political activities	1.12	1.54 (107.31)	-6.542**
Online civic activities	1.74	2.00 (317)	-3.88**
Size of political discussion network	5.78	14.50 (90.58)	-4.13**
Other people have same knowledge	0.92	0.90 (125.37)	2.16*

* $p < .05$, ** $p < .01$

Political bloggers were also more likely than personal bloggers to have reported a specific issue when asked ‘What do you feel is the most important issue facing the local area?’ and to report that they would attend a public meeting on that issue. They were considerably more likely to speak at such a forum and to express an opinion that is different from others at the forum. Finally, political bloggers were more likely to report the Internet helped them to become more involved in local issues. See Table 9.

Table 9: Political vs Personal Bloggers on Important Issue

Variable Labels	Mean		<i>t</i> test (df)
	Personal	Political	
Important issue reported or not	0.84	0.94 (256.12)	-2.83**
Likely to attend public forum on issue	2.62	3.40 (219.41)	-6.82**
Likely to speak at public forum	2.75	3.41 (206.87)	-5.45**
Likely to express different opinion	3.22	3.62 (181.12)	-4.22**
Internet helpful for involvement	2.29	2.54 (316)	-2.40*

* $p < .05$, ** $p < .01$

Finally, bloggers who were politically active reported significantly more political discussions in the local groups with which they are involved than did bloggers who were not politically active. This provides further evidence of the opportunities for political discussion and influence engaged in by opinion leaders with members of their social network.

5. Discussion

While our research was conducted using households in Blacksburg and Montgomery County, Virginia, this ICT rich region may indicate trends that may also be observed elsewhere. Clearly, our respondents' uses of information and communication technology, particularly such collaborative tools as blogs, shapes their political attitudes, efficacy and behavior. The present findings suggest that people with higher levels of political participation use information technology in a variety of ways, including many traditional ways (e.g., email government officials, staying informed). Even so, one of the more innovative forms of Internet use, blogging, by politically and socially active respondents may foster more political engagement and deliberation because they act as opinion leaders influencing members of their social network.

Respondents who are more frequent blog readers, commentators and/or writers tend to be younger, male, and affiliated with more informal groups than those who have never heard of blogs. Respondents who are *politically active* and who also read, comment on and/or write blogs are slightly older (average age 47), more extroverted, and are affiliated with more local formal groups than those who are *not politically active* and read, comment on and/or write blogs. 'Political' bloggers have more income and larger households than do 'personal' bloggers; 'political' bloggers also have a higher sense of community collective efficacy (the belief that a community can solve collective problems despite obstacles) than do personal bloggers. Finally, political bloggers are more likely than personal bloggers to report a problem in the local area that they consider to be important. They are more likely to: 1) have talked with people about the issue, 2) attend a public forum about it, 3) speak at such a forum, and 4) express an opinion that differs from others at a public forum on the issue. They discuss the issue with significantly more people and with people who have different levels of knowledge on issues (probably less) than themselves.

These findings reinforce theoretical expectations that these politically and socially active people are opinion leaders and that they are beginning to use innovative information and communication tools to support their political interests and influence among their social circles. The highly social nature of these persons' political behavior (i.e., more discussion with larger social networks) could easily migrate into the realm of blogging, given the informal and conversational nature of blogs. Presently, we lack specific questions in our survey regarding the type of blogs respondents are reading, commenting on, or writing. This is a limitation of our study. We are pursuing this line of questioning in a series of focus group interviews with our respondents in our ongoing research.

We expect that tools such as blogs and wikis, because they allow people to organize and collaborate among themselves, will foster citizen-to-citizen discussion and deliberation, led by influentials. The conversational style of blogs, combined with their easy accessibility, suggests that Internet users will continue to adopt

them at increasing rates. It is also likely that blogs and wikis will increasingly be used to express citizens' views and strengthen connections with existent social networks. Given greater ease and possibilities for informal exchange of information and ideas among friends, family members, and other citizens—these new self-organizing, collaborative groups may well help to enhance deliberative and engaged political discourse within democratic societies.

6. ACKNOWLEDGMENTS

We are grateful for support of this research from the National Science Foundation Digital Government Program (IIS-0429274). We would also like to thank our collaborators Daniel Dunlap and Mary Beth Rosson, with special thanks for research assistance from B. Joon Kim, Jaideep Godara, Alain Fabian, William Randolph and Andrew Mike.

7. REFERENCES

- [1] Arterton, F.C. *Teledemocracy: Can Technology Protect Democracy?* Sage, Newbury Park, CA, 1987.
- [2] Bandura, A. *Self-efficacy: The Exercise of Control*. Freeman, New York, 1997.
- [3] Barber, B. *Strong Democracy: Participatory Politics for a New Age*. University of California Press, Berkeley, CA, 1984.
- [4] Carroll, J.M. and Reese, D. Community collective efficacy: Structure and consequences of perceived capacities in the Blacksburg Electronic Village. *Hawaii International Conference on System Sciences, HICSS-36* (January 6-9, Kona) 2003.
- [5] Carroll, J.M., Rosson, M.B., Dunlap, D., Kavanaugh, A., Schafer, W. and Snook, J. Social and civic participation in a community network. In R. Kraut, M. Brynin and S. Kiesler (eds.) *Domesticating Information Technologies*. Oxford University Press, New York, 2005.
- [6] Coleman, S. and Gotz, J. *Bowling Together: Online Public Engagement in Policy Deliberation*. 2002. Downloaded from: <http://bowlingtogether.net/>
- [7] Dahl, R. *Democracy and its Critics*. Yale University Press, New Haven, CT, 1991.
- [8] Dahlberg, L. The Internet and democratic discourse: Exploring the prospects of online deliberative forums extending the public sphere. *Information, Communication & Society* 4, 4 (2001), 615-633.
- [9] Fishkin, J.S. *Democracy and Deliberation*. Yale University Press, New Haven, CT, 1991.
- [10] Gastil, J. and Levine, P. (eds.) *Deliberative Democracy Handbook: Strategies for Effective Civic Engagement in the 21st Century*. Jossey-Bass, San Francisco, CA, 2005.
- [11] Horrigan, J. *Online communities: Networks that Nurture Long-Distance Relationships and Local Ties*. Pew Internet & American Life Project, 2001. <http://www.pewinternet.org>
- [12] Horrigan, J., Garrett, K., and Resnick, P. *The Internet and Democratic Debate*. Pew Internet & American Life Project, 2004. <http://www.pewinternet.org>

- [13] Jacobs, J. *The Death and Life of Great American Cities*. Random House, New York, 1961.
- [14] Katz, E. Communications research since Lazarsfeld. *Public Opinion Quarterly* 51 (1987), 525-545.
- [15] Katz, E. and Lazarsfeld, P. *Personal Influence: The Part Played by People in the Flow of Mass Communications*. The Free Press, New York, 1955.
- [16] Katz, J. and Rice, R. *Social consequences of Internet use*. MIT Press, Cambridge, MA, 2002.
- [17] Kavanaugh, A., Cohill, A. and Patterson, S. The use and impact of the Blacksburg Electronic Village. In A. Cohill and A. Kavanaugh (eds.), *Community Networks: Lessons from Blacksburg, Virginia*. Artech House, Norwood, MA, 2000, 77-98.
- [18] Kavanaugh, A., Reese, D.D., Carroll, J.M., and Rosson, M.B. 2003. Weak ties in networked communities, pp. 265-286. In M. Huysman, E. Wenger and V. Wulf (eds). *Communities and Technologies*. Kluwer Academic Publishers, The Netherlands. Reprinted in *The Information Society* 21, 2 (2005), 119-131.
- [19] Kavanaugh, A., Carroll, J.M., Rosson, M.B., and Zin, T.T. Participating in civil society: The case of networked communities. *Interacting with Computers* 17 (2005a), 9-33.
- [20] Kavanaugh, A., Isenhour, P., Cooper, M., Carroll, J.M., Rosson, M.B., and Schmitz, J. Information technology in support of public deliberation In P. Besselaar, G. de Michelis, J. Preece, and C. Simone (eds.) *Communities and Technologies 2005*. Springer, The Netherlands, 2005b, 19-40.
- [21] Kavanaugh, A., Isenhour, P., Godara, J., Cooper, M., Midha, A., and Randolph, W. Detecting and Facilitating Deliberation at the Local Level. In T. Davies and B. Noveck (eds.) *Online Deliberation: Design, Research and Practice*. Chicago, IL: University of Chicago Press, Forthcoming.
- [22] Keller, E. and Berry, J. *The Influentials*. The Free Press, New York, 2003.
- [23] Kim, J., Wyatt, R. and Katz, E. News, talk, opinion, participation: the part played by conversation in deliberative democracy. *Political Communication* 16, 4 (1999), 361-385.
- [24] Kirn, K. Building social capital on the web: The case of Minnesota E-Democracy. In Turow, J (Ed.), *Energizing Voters Online: Best Practices from Election 2000*. Report No. 39, Annenberg Public Policy Center, University of Pennsylvania, Philadelphia, PA, 2002.
- [25] Kraut, R., Scherlis, W., Mukhopadhyay, T., Manning, J., and Kiesler, S. The HomeNet field trial of residential Internet services, *Communications of the ACM*, 39 (1996), 55-63.
- [26] Kraut, R., Kiesler, S., Bonka, B., Cummings, J., Helgelson, V., and Crawford, A. Internet paradox revisited, *Journal of Social Issues*, 58 (2002), 49-74.
- [27] Krugman, P. *The Self-Organizing Economy*. Blackwell Publishers, Oxford, UK, 1996.
- [28] Lave, J. and Wenger, E. *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press, Cambridge, UK, 1990.
- [29] Markus, M.L. Toward a “critical mass” theory of interactive media: Universal access, interdependence and diffusion. *Communication Research*, 14, 5 (1987), 491-511.
- [30] Michaelson, M. Political efficacy and electoral participation of Chicago Latinos. *Social Science Quarterly*, 81, 1 (March 2000), 136-150.
- [31] Milbrath, L. and Goel, M. *Political Participation: Why and How Do People Get Involved in Politics?* University Press of America, Lanham, MD, 1977.
- [32] Miller, A., Goldberg, E., and Erbring, L. Type-set politics: participation, representation , and policy preferences. *American Political Science Review* 73, 1 (1980), 67-84.
- [33] Nardi, B. Why we blog. *Communications of the ACM*, 47, 12 (2004), 41-46.
- [34] Norris, P. 2001. *Digital divide: Civic engagement, information poverty and the Internet*. New York: Cambridge University Press.
- [35] Price, V. 2005. Online Health Discussion Project. Paper presented at Stanford Online Deliberation conference, May 19-21, 2005.
- [36] Rainie, L. 2005. *The State of Blogging*. Norris, P. *Digital Divide: Civic Engagement, Information Poverty and the Internet*. Cambridge University Press, New York, NY, 2001.
- [37] Price, V. *Online Health Discussion Project*. Paper presented at Stanford Online Deliberation Conference, May 20-22, 2005.
- [38] Rainie, L. 2005. *The State of Blogging*. Pew Internet & American Life Project, <http://www.pewinternet.org>
- [39] Rogers, E.M. *Communication Technology: The New Media in Society*. Free Press, New York, NY, 1986.
- [40] Rogers, E. and Shoemaker, F. *Communication of Innovations* (2nd edition), The Free Press, New York, 1971.
- [41] Schmitz, J., Rogers, E., Phillips, K., and Paschal, D. The Public Electronic Network (PEN) and homeless in Santa Monica. *Journal of Applied Communication Research* 23, 1 (1995), 26-43.
- [42] Schudson, M. The limits of teledemocracy. *The American Prospect*. (Fall, 1992), 41-45.
- [43] Schudson, M. Why conversation is not the soul of democracy. *Critical Studies in Mass Communication*, 14 (1997), 297-309.
- [44] Schuler, D. *New Community Networks: Wired for Change*. ACM Press, New York, NY, 1996.
- [45] Verba, S. and Nie, N. *Participation in America: Political Democracy and Social Equality*. Harper and Rowe, New York, NY, 1972.
- [46] Wheatley, M.J. *Leadership and the New Science*. Berrett-Koehler, San Francisco, CA, 1992.
- [47] Wiley, D. and Edwards, E. *Online Self-Organizing Social Systems: The Decentralized Future of Online Learning*. 2003. Downloaded from: <http://wiley.cc.usu.edu/>
- [48] Wulf, V. Evolving cooperation when introducing groupware: A self-organization perspective. *Cybernetics and Human Knowing*, 6, 2 (1999), 55-75.

ⁱ See also: The Berkman Center for Internet & Society (<http://cyber.law.harvard.edu/projects/deliberation>); The Civic Exchange Strong Democracy in Cyberspace (<http://webserver.law.yale.edu/infosociety/civicexchange.html>); Deliberative Democracy Consortium (<http://deliberative-democracy.net>); National Science Foundation Digital Government Project (<http://digitalgovernment.org>); National Coalition for Deliberative Democracy (<http://www.ncdd.org>); Minnesota E-Democracy (<http://www.e-democracy.org>); among others.

ⁱⁱ Notes on Response Codes for Table 1 & Table 2:

^a 1=Town of Blacksburg, 2=Town of Christiansburg, 3= Montgomery County.

^b 1= less than \$ 50,000, 2= \$ 50,000 or more.

^c 1=eighth grade or less, 2=some high school, 3= high school grad GED, 4=some college/ certificate program, 5= graduated from college or certificate program, 6=some graduate level work, 7=completed graduate school/ professional school.

^d 1= married, 2=single, 3=divorced, 4=separated, 5=widowed, 6= living with partner.

^e 1= some of the HH members are younger than 18 years, 0= none of the HH members are younger than 18 years.

ⁱⁱⁱ Notes for Tables 7-9:

Number of respondents who never heard of blogs ranged from 233 to 236.

Number of respondents who heard of blogs ranged from 317 to 319.

Degrees of freedom (*df*) with decimal numbers are from equal variance NOT assumed t tests. *df* with whole numbers are from equal variance assumed *t* tests.

Bringing an Informed Public into Policy Debates through Online Deliberation: The *Healthcare Dialogue* project

Vincent Price

Annenberg School for Communication
University of Pennsylvania
Philadelphia, PA
(215) 573-1963

vprice@asc.upenn.edu

Joseph N. Cappella

Annenberg School for Communication
University of Pennsylvania
Philadelphia, PA
(215) 898-7059

jcappella@asc.upenn.edu

ABSTRACT

A year-long experiment was conducted to better understand the potential of web-based deliberations to inform public policy. Focused on health care reform, the project drew from periodic surveys and a series of online group deliberations to examine the interaction of policy elites and ordinary citizens in online settings, and to test hypotheses related to group composition, discussion processes, and decision making. This paper describes the project design and summarizes several key findings to date.

Categories and Subject Descriptors

H.5.3 [Information Systems]: Group and Organization Interfaces – *computer-supported cooperative work, evaluation/methodology, synchronous interaction, web-based interaction*.

General Terms

Measurement, Design, Reliability, Human Factors, Theory.

Keywords

Digital Democracy, Online Deliberation, Health Care Policy.

1. INTRODUCTION

Many models of policy making consider direct popular engagement in determining policy as being of dubious value, particularly when the issues at stake are complex. Instead these models rely on an unrepresentative layer of experts to define the public's welfare [2]. However, a number of scholars have called on the policy sciences to include lay citizens directly in technical policy deliberations, arguing that ordinary citizens, when engaged in issue-based deliberations, will form more considerate and informed opinions, worthy of guiding difficult policy options [2,3]. Internet technologies now permit deliberations among geographically dispersed groups, potentially bringing far greater

reach and increased representation to deliberative democracy [1].

The *Healthcare Dialogue* project examines whether web-based deliberation among geographically dispersed and informationally diverse persons can lead to more informed public opinion on health care issues, despite their complexity. Small-group, online policy discussions among health-care elites and ordinary citizens alike were experimentally constructed and studied. Project objectives include: (a) examining online deliberation as a means of maximizing public influence in policy making; (b) studying the interaction of policy elites and ordinary citizens in online discussions; and (c) testing hypotheses related to group composition and the quality of deliberations and outcomes.

2. DESIGN

The research involved a multi-group, multi-wave panel design, beginning with a baseline survey conducted in the summer of 2004 ($N=2,497$). A random subset of participants engaged in a series of online discussions about health care issues facing the country in the fall of 2004 (eighty discussion groups meeting twice each) and again in the late winter and early spring of 2005 (fifty groups meeting twice each). The four discussion waves took place in September and November 2004 and in February and April 2005, with each consisting of a brief pre-discussion survey followed by an hour-long online chat followed by another brief post-discussion survey. An end-of-project survey of participants was conducted in August 2005 (completed by roughly three-quarters of baseline respondents, $N=1,830$).

2.1 Sample

The project employed a stratified sampling strategy, such that the final baseline sample represented both a general population sample of adult citizens, age 18 or older ($N=2,183$), as well as a purposive sample of health care policy elites with special experience, knowledge, and influence in the domain of health care policy and reform ($N=314$). The general population sample was further stratified into members of “issue publics” who are highly attentive to and knowledgeable about health care issues ($N=804$), and ordinary citizens ($N=1,379$). Comparisons of the obtained, unweighted baseline general population sample to a representative telephone (RDD) sample fielded on the same days and the U.S. Census data indicate that the samples are broadly comparable, although project participants are somewhat more likely to be middle aged and to follow politics more frequently.

2.2 Experimental design

A subset of the baseline panel (262 health care policy elites; 461 issue-public members; 768 ordinary citizens) was randomly assigned within strata to participate in the four moderated online group discussions, including pre- and post-discussion surveys, which were conducted over the course of the year. Forty of the discussion groups were designed to be homogenous within strata (8 elite only, 12 issue-public only, 20 general citizens only); the other forty were mixed across strata. A control group was asked to complete project surveys but not invited to the discussion meetings (52 policy elites; 343 issue-public members; 611 ordinary citizens).

Because baseline surveys indicated broad agreement that the most pressing problems facing the health care system included the rising costs of health insurance, the large number of uninsured Americans, and the rising costs of prescription drugs, these issues became the focus of the online deliberations. Eighty groups (8 homogeneous elite; 12 homogeneous issue-public; 20 homogeneous general citizen; 40 heterogeneous across strata) met twice in the fall of 2004 to discuss insurance-related issues. A total of 614 project participants (123 elites; 206 issue-public members; 285 general citizens) attended at least one of the two discussions. The subset of 614 fall discussion attendees was then reassigned to 50 new groups for another round of two discussions in the spring of 2005, focusing on prescription drugs. In this second round of online deliberations, half the participants remained in homogeneous or heterogeneous groups as before, while half were switched (from homogeneous to heterogeneous groups, or *vice versa*).

3. FINDINGS TO DATE

3.1 The Health Care Issue Public

Contours of the issue public for health care policy have been examined extensively using project survey data [4]. Its cognitive, affective, and behavioral underpinnings (measured through health care knowledge, holding strong opinions on health care issues, and participation in health-related political activities) have been shown to be only weakly interconnected; hence a multi-dimensional index combining all three factors was developed for identifying issue-public members.

3.2 Online Discussions

Within the assigned discussion panel, online group attendees tended to be significantly older, better educated, more knowledgeable, more engaged in health care issues, and more likely to be white than non-attendees. No significant differences were observed, however, in gender or political leanings. Use of an asynchronous bulletin board, encouraged among a subset of control participants, was far less utilized than the synchronous group meetings. A comprehensive, multivariate model of propensity to participate in the online discussions has been developed for use in balancing the experimental and control conditions [5].

Impressions of the online group deliberations were very positive, with most attendees indicating that they found the discussions interesting. General population participants rated the experience significantly more positively than did health care elites. Groups across all strata expressed high levels of satisfaction with their

final choice of a top-priority policy proposal for addressing health insurance (elite $M=3.9$ on a 5-point scale; issue-public $M=3.9$; general citizen $M=4.13$).

3.3 Effects of Deliberation

Analysis of the first phase of discussions provided evidence of significant deliberation effects [5]. Results suggest (a) that participation in online deliberations leads to higher levels of opinion holding on matters of health care policy; (b) that participation leads to substantive and interpretable shifts in policy preferences; and (c) that the shifts induced by deliberation reflect movement toward more informed and politically sophisticated positions. More specifically, after controls for propensity to attend, preferences at baseline, and other background characteristics, attendees were less likely than non-attendees to support tax-based reforms and were more supportive than non-attendees of government programming and regulations as a means to cut insurance costs. These differences between participants and non-participants parallel those between elites and general citizens at baseline; however, they occurred to a greater degree in groups *without* elite members, suggesting they were not the mere product of elite persuasion. [5].

4. CONTRIBUTIONS

The research aims to make significant theoretical contributions to understanding elite/mass relationships in a democratic society, and to lend practical guidance for designing deliberative encounters in service of public policy. Better understanding the barriers to effective conversations across social groups and within the on-line environment will permit regulatory groups and legislative bodies to involve citizens in fruitful deliberations.

5. ACKNOWLEDGMENTS

This research is supported by grants from the Annenberg Public Policy Center of the University of Pennsylvania and the National Science Foundation (Grant EIA-0306801). Views expressed are those of the authors and do not necessarily reflect those of the sponsoring agencies.

6. REFERENCES

- [1] Becker, T., & Daryl Slaton, C. *The future of teledemocracy*. Preager, Wesport, CT, 2000.
- [2] deLeon, P. Democratic values and the policy sciences. *American Journal of Political Science*, 39 (1995), 886-905.
- [3] Fischer, F. Citizen participation and the democratization of policy expertise: From theoretic inquiry to practical cases. *Policy Sciences*, 26 (1993), 165-88.
- [4] Price, V., David, C., Goldthorpe, B., McCoy Roth, M., & Cappella, J.N. Locating the issue public: The multi-dimensional nature of engagement with health care reform. *Political Behavior*, in press.
- [5] Price, V., Feldman, L., Freres, D., Cappella, J.N., & Zhang, W. *Informing public opinion about health care reform through online deliberation*. Annual meeting of the International Communication Association, Dresden, Germany, May, 2006.

Decoding Political Discourse Networks

John Kelly

Columbia University
2950 Broadway, MC
New York, NY 10027
+1 646 485 7364
kjwl@columbia.edu

David Stark

Columbia University
1180 Amsterdam Ave
New York, NY 10027
+1-212-854-3972
dcs36@columbia.edu

ABSTRACT

This describes the current research project on political discourse networks, at the Center on Organizational Innovation at Columbia University. This research is supported by the NSF grant #IIS- 0441999, Technologies of Civil Society.

Keywords

Public sphere, democracy, e-government, politics, online forum.

1. INTRODUCTION

This research is concerned with the role new technologies are playing in the evolution of our “public sphere.” Discussion forums, websites, listservs, blogs, and other modes of online political communication are often said to be changing the dynamics of democracy. Different technologies, e.g. blogs and threaded discussion, seem to have different implications, arising from the interplay of technological affordances, user behavior, and the political culture of the society at large. New methods are required to analyze and understand emerging networks of online political interaction, and to assess their characteristics in light of normative democratic theory.

2. CURRENT PROJECT ACTIVITIES

2.1 Overview

Though most online political discussion is “public” in the sense of being visible to anyone with internet access, techniques for analyzing these discussions and models for characterizing the forms of online discourse are at best rudimentary. Seeking to advance the methodology for analyzing online discourse, our approach combines content analysis with network analysis, to build models of political discourse networks among actors identified by their ideological characteristics and issue positions. This innovative combination of methods is used to look at the relationship between discursive style and political attitudes on the one hand, and network structure and actor position on the other. The patterns of association among these factors reveal how ostensibly anarchic online discourse communities are structured by group norms and participant interests.

2.2 Conceptual Background

Do online political discussions tend to aggregate diverse voices in cross-cutting debate and deliberation? Or do “audiences” for online discussion tend to fragment into

ideological echo chambers? Online networks of political discourse emerge from billions of individual choices by millions of individual citizens about what to discuss online, where to discuss it, and with whom. Do these choices lead individuals to interact across ideological divides, or to cluster within them?

It is possible that individual-level selectivity biases, enabled by the internet’s facilitation of choice in information consumption and mediated interaction, will precipitate a balkanization of political discourse. But though the tendency toward political homophily in social networks is well established, we should not assume it necessarily operates in online discussion environments. Some online discussants seek reinforcement, but others go online to encounter differing points of view. Individual motivations vary, and therefore so do individual behaviors, and ultimately the structures of discussion networks that emerge from them. If debate is the modal activity in a discussion environment, it is conceivable that online fora are enabling exactly the kind of public commons the internet is often said to endanger.

2.3 Methods

Our approach combines content analysis and network analysis. A corpus of threaded discussion is selected, based on criteria relevant to theories of political communication (e.g. differences between *attentive* and *issue* publics). Then, content analysis is used to determine ideological characteristics of authors, their positions on particular issues, and aspects of their discursive behavior. Software is then used to examine large-scale patterns of interaction among both coded and additional non-coded authors, capturing behavior in interactions beyond those in the initial sample corpus used for content analysis. This approach is focused on leveraging the value of labor intensive, qualitative techniques applied to dozens of authors in hundreds of interactions against the power of advanced computer techniques to analyze thousands of authors in millions of interactions.

3. COLLABORATORS

Our work has been pursued in fruitful collaboration with the Community Technologies Group at Microsoft Research, and with Warren Sack at UC Santa Cruz. Prof. Sack has developed advanced techniques for analyzing “very large-scale conversations” online. MSR has developed Netscan, an excellent tool that has captured, and provides advanced capabilities for analyzing, nearly all threaded discussion in USENET for the last six years. Our MSR collaborators are pioneers in the quantitative analysis of online communities,

focused broadly on network structures that arise from various types of online discussion groups, such as technical, support, and fan. Our principal MSR collaborators are:

Danyel Fisher, Researcher, Community Technologies Group, Microsoft Research

Marc Smith, Director, Community Technologies Group, Microsoft Research

4. RESEARCH CONTRIBUTIONS

In addition to conference presentations at Stanford's Online Deliberation conference (May 2005), three papers are currently awaiting publication:

1. "Debate, Division, and Diversity: Political Discourse Networks in USENET Newsgroups" John Kelly, Danyel Fisher, Marc Smith. Submitted, *Political Communication*. [also: <http://www.coi.columbia.edu/workingpapers.html>]
2. "Searching the Net for Differences of Opinion" John Kelly, Warren Sack, and Michael Dale. Publication forthcoming, as book chapter in *Online Deliberation* (University of Chicago Press, 2006) [online at: <http://www.coi.columbia.edu/workingpapers.html>]
3. "Friends, Foes, and Fringe: Visualizing Structure and Position in Political Discussion Networks." John Kelly, Danyel Fisher, Marc Smith. Publication forthcoming, as book chapter in *Online Deliberation* (University of Chicago Press, 2006)

5. EXAMPLES

Following are two figures illustrating research results.

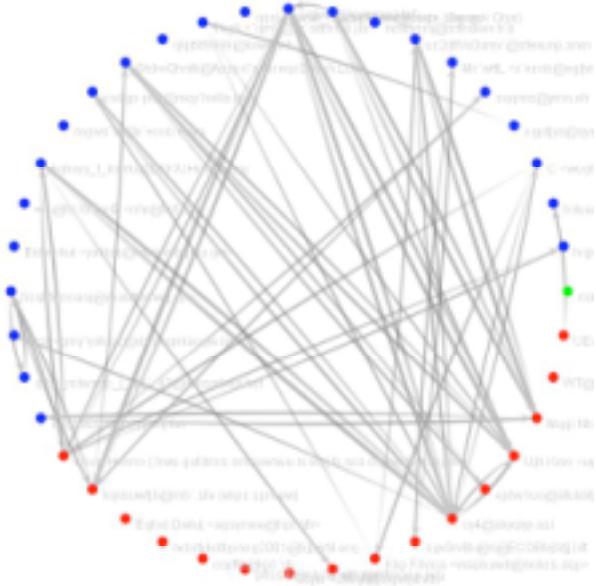


Figure 1: author interactions and political clusters

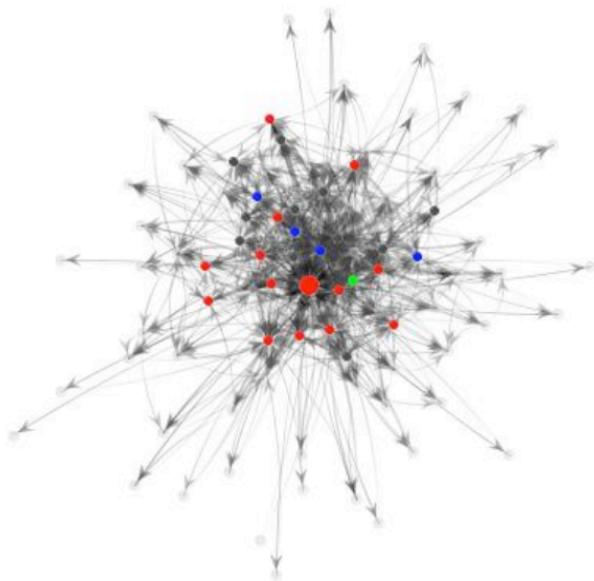


Figure 2: ego-network within coded political group

6. NEXT STEPS

Reaction to this research, at the Stanford Online Deliberation 2005 conference and elsewhere, has been very positive. There is widespread awareness that the dynamics of rapidly emerging online discourse communities must be made tractable to communications research. This work is generally received as making a contribution in that direction.

Challenges and opportunities exist for extending this work. We have discovered some very strong principles of organization of threaded political discussion, e.g. that cross-cluster debate is the modal activity. But there are clearly other principles at work which must be explored, e.g. the role of "discourse quality" in aggregating discussants.

A key challenge is adapting the techniques developed for analyzing threaded discussion in USENET to other discussion environments and technologies. Of key interest are:

1. Discussion environments that allow user feedback to bias the subsequent presentation of discussion posts (e.g. Slashdot);
2. Environments that do not support the overt linking of one message to another, such as the "talkback" features of many news sites allowing reader comments on specific articles and news reports;
3. Feedback on blogs, which generally aggregate the politically like-minded but occasionally feature differences of opinion.

Advancing our understanding of online politics in these directions can benefit scholars, policymakers, and technology designers alike, helping inform a discussion about the implications of online discourse in light of normative values of democracy.

SESSION 3C

EMERGENT DESIGN

Moderator

Alan Borning, University of Washington, USA

Titles and Authors

Water Models and Water Politics: Design, Deliberation, and Virtual Accountability
Jackson, Steven

Organic Development: A Top-Down and Bottom-Up Approach to Design of Public Sector Information Systems
Tyworth, Michael; Sawyer, Steve

Water Models and Water Politics: Design, Deliberation, and Virtual Accountability

Steven Jackson

School of Information

University of Michigan

301D West Hall, 1085 S University

Ave

Ann Arbor, MI 48109-1107

734-764-8058

sjackso@umich.edu

ABSTRACT

Computer simulation models have emerged in recent decades as increasingly prominent technologies within the toolkit of modern democratic governance. Despite and/or because of this centrality, however, formerly ‘technical’ domains of modeling have been opened up to new forms of public debate, scrutiny and critique, with uncertain policy consequences. This paper traces such dynamics through one field of contemporary relevance: the joint evolution of simulation models and water management in California. Rather than decrying the politicization or debasement of expertise, I argue that broadening the deliberative basis of model design and use is likely to improve both the technical and political functioning of models. The paper concludes by sketching a model of ‘virtual accountability’ meant to inform the actions of future model builders, users, and stakeholders in contested realms of public policy.

General Terms

Management, Design, Standardization, Theory, Verification.

Keywords

Simulation, Validation, Models, Water, Environment, Policy, Governance, Infrastructure.

1. INTRODUCTION

Numeric models are as old or older than the practice of modern science and government themselves. They have long played a central role at the three-way interface of theory, data, and action, generating the sorts of hypotheses, predictions, and proofs that have become staple elements within liberal democratic regimes of justification and public action [13, 34]. With the emergence of digital computing, the role of models in science and administrative decision-making has taken a qualitative step forward. From physics and the bomb, computer simulation techniques have spread rapidly through science, industry, and public policy in the post-war period, most notably in well-funded

domains where theoretical, observational, and experimental roads to knowledge have proven impractical, whether for reasons of cost, accessibility, or ethical sensitivity. These have included the big (e.g. climate science), the small (e.g. genetic research), the remote (e.g. astrophysics, mineral exploration), the vastly distributed (e.g. epidemiology, economics), and the humanly, ecologically, or politically fragile (e.g. medical research, post-proliferation nuclear weapons testing) [12, 16, 17, 23, 31, 32].

Despite this remarkable growth, there has been as yet little work addressing simulation models as *policy technologies*, i.e. artifacts whose technical shape, deliberative character, and public effectiveness are built and adapted through mutually iterative processes of technical and political refinement. To understand this dynamic, digital government scholarship must find answers to a number of prior questions. What role have models come to play in mediating the deep political and epistemic tensions characterizing such fields as environmental management and policy? What barriers and limitations have models encountered in their movement from academic and agency science into the world of public deliberation and decision-making? How have the distinctive characteristics of that world acted back upon the technical work of simulation, influencing the design histories and affordances of the models themselves? Finally, what prescriptive lessons can be drawn from these experiences to guide the future work of model builders, users, and stakeholders in contested realms of public policy?

Drawing on scholarship in Science and Technology Studies (STS) and the methods of the qualitative social sciences, this paper addresses such questions through the medium of a single case study: the technical and political evolution of water simulation models in California and the American Southwest, and their shifting place within the region’s high-stakes and often fractious water management regime.¹

¹ Principal fieldwork for this paper was carried out between January 2003 and July 2005, and included ethnographic observation and more than 60 interviews with modelers, policymakers, and environmental, urban, and agricultural representatives. For a more complete description of study methodology and findings, see [18].

2. MODELS IN ECOSYSTEM SCIENCE AND POLICY

Among the policy fields in which computer modeling has assumed a prominent post-war role, environmental science and management have represented something of a perfect storm: they have tended to draw on vast amounts of data sourced from highly distributed collection regimes; they deal regularly in large, disparate, and complexly interconnected data sets; they occupy research fields in which the possibilities of experiment or direct observation are prescribed by constraints of time, scale, safety, or physical access; and they have grown in partial response to social movements and agencies with a strong pragmatic interest in prediction [12, 30]. Under such circumstances, models have come to play significant and indeed multiple roles: from the shaping of observation and experiment, to the synthesis of large and disparate data sets, to the generation of numeric predictions and forecasts [6]. In each case, model builders and users have faced significant challenges. Modeling work has frequently evolved in relative isolation from other analytic and fieldwork traditions, leaving key model parameters, data sets, and functional relationships incomplete, ill-documented, unevenly maintained, or otherwise underspecified [2, 10, 24]. Modelers in many fields have struggled to define appropriate relationships between data sets of radically varying types and quality, especially where real-world functional dynamics – often the question of central analytic concern – remain poorly understood. Additional theoretical and practical challenges greet efforts to define the appropriate scope and granularity of simulation: how much of the (impossibly vast) world ought to be accounted for within the confines of the model, and what can be safely left outside of it? At what point does the radical and fine-grained inclusion of variables obscure the analytic clarity, general comprehensibility, and pragmatic workability of models? Here, modelers have debated the relative merits of ‘fast-and-frugal’ models (marked by speed, agility, access, and rapid evolution) versus their larger, slower, and more computationally intensive relatives [7].

Modelers have adopted a variety of strategies for dealing with such uncertainties. A first and obvious response has been to measure model results against real world experience, adducing direct or indirect evidence for and against the reliability of model results. But this check against external record, simple in concept, may be surprisingly difficult to accomplish in practice – not least because models are frequently deployed precisely where possibilities of observational or experimental knowledge are at their weakest. Under such circumstances, ‘crucial tests’ able to conclusively confirm or disconfirm model results are in remarkably short supply.

In response, model builders and users have turned to a variety of second-order strategies to support the validity of their claims. In sensitivity analyses, modelers tweak variables in a sequential manner, measuring the degree of responsiveness or “sensitivity” registered across key parameters of the system. Sensitivities deemed to be inordinate when measured in this way may indicate the presence of artifactual properties likely to skew model results vis-à-vis the real-world systems under study. A second response to uncertainty in simulation has been the historical calibration or validation exercise, where the model is evaluated according to its ability to reproduce known results captured from historical observation. Under the simplest description of this, the model is

loaded with the initial conditions and known input data for a given period, run, and the results compared to the historical record. Alternatively, the period of calibration may be split into two separate phases: an initial ‘tuning’ period, in which model parameters are adjusted until a good fit with historical data is achieved; and a second testing phase, in which the corrected model is run and compared against the remainder of the record. Models capable of reproducing known history with reasonable accuracy are inferred to reliably mimic and project the performance of the system into the future. Models tested in this way are frequently said to be “verified” or “validated.”

Such second-order responses to uncertainty face problems of their own, however. As Oreskes et. al. point out, in contrast to the possibility of closure existing in formal logic and mathematical systems, the ‘systems’ of ecosystem science and management remain radically and inevitably open: input parameters are incompletely known, model elements remain subject to uncertain scale effects, available data is of uneven quality and coverage, and the fit of model to world depends on a prior set of ultimately untested and unmodeled assumptions [1, 27]. In practice, efforts at historical validation as described above are often in a strict sense inconclusive, since model tuning constitutes an ordinary and ongoing part of model life, limiting the degree of autonomous verification that may be said to follow from successful replication of the historical record. Even where the historical match is good, the frequent non-linear character of earth systems, when combined with the generally short periods and places for which good historical data is available (as compared to the scales of time and space at which models are often called upon to predict) raises the problem that currently negligible mismatches between model and world may lead to significant divergences over time: models only marginally wrong over the recent past may prove to be significantly, even catastrophically, wrong when projected far enough into the future. Moreover, successful calibration does not necessarily imply a ‘true’ or ‘realistic’ understanding of underlying causal mechanisms and dynamic; models can produce the ‘right’ numbers for the wrong reasons, and the significance of fundamental conceptual error may increase as the system moves downstream in time. Beyond all this, the fundamental openness of ecological systems means there is no guarantee against future changes in the system that might render models reasonably accurate in replicating past and present observational results wildly inaccurate when projected into the future. Under such circumstances, as Oreskes and Belitz observe, efforts to train simulations to historical data may exacerbate the conservative bias of models by extending, sometimes without justification, current trends into the future. Similar dynamics govern the general exclusion of low probability events, whose collective impact may further skew model results [28].

Such narrowly ‘technical’ challenges take on new life and complexity as they travel beyond the immediate boundaries of professional communities of practice. An important impetus for the development of specific models and the growth of simulation technique more generally has come from agencies and policy-makers – themselves driven by changes in the political and regulatory field wrought in large part by various environmental movements – who have sought in models new and authoritative bases on which to ground regulatory and policy actions. In this regard, modelers work in a field simultaneously constrained and constituted through the presence of what historian of science Ted

Porter has described as ‘powerful outsiders’: figures from beyond the immediately technical realm who nevertheless exert a significant influence over the shape and form of its internal deliberations and practices [29]. One effect of this positioning vis-à-vis the decision requirements of the policy realm may be to push models in the direction of a steadily harder predictive stance, as sharply bounded simulation results pass into the hands of actors for whom technical caveats, uncertainties, and other limitations will remain at least opaque and perhaps an impediment to efficient decision-making. Under such circumstances, predictions may take on an aura of finality that makes even their producers uneasy. The fallibility or fragility of model knowledge may disappear in translation, peeled away on the margins of the science-policy interface. By virtue of such dynamics, model predictions, like other scientific findings, will frequently appear to ‘harden’ as they travel outwards; in Harry Collins memorable phrase: “distance lends enchantment” [9]. Where such dynamics are most fully developed (e.g. the climate change debates), this has tended to produce an all-or-nothing public response to model credibility: models are too frequently regarded as all right or all wrong, with little room for a nuanced and ultimately more helpful engagement with model strengths and weaknesses.

The upshot of all this is that the degree of certainty both professionally and popularly assigned to model predictions may be systematically distorted. As Oreskes et. al. note, routine terms of art such as ‘verification’ and ‘validation’ may mislead, particularly where restricted professional usages meet the more expansive and unqualified meanings generally assigned these terms in common discourse [27]. Under such conditions, assessments of fit and adequacy remain deeply situated exercises. Competent and accredited professionals may well (and frequently do) arrive at different conclusions as to the validity and appropriateness of a given model application, even where using the same data. External stakeholders (e.g. funders, government agencies, scientific review panels, industry officials, public interest groups, etc.) will bring criteria of fit all their own, which may contradict, but also significantly shape, the internal deliberations of model builders and users. To the extent that models (and model results) function as ‘boundary objects’ shared between multiple social and institutional worlds, assessing their adequacy is an inescapably practical as well as narrowly technical affair [31, 33]. To the extent that models are purposeful, i.e. embedding particular predilections or orientations to action, questions of design are inextricably bound with questions of application. In this context, the central question of model evaluation is not ‘is it good?’ but rather “is it good enough for the purpose?” [15].

Under such circumstances, the contributions of modeling to wider arenas of public debate should be offered and received in a spirit of responsible and critical humility [19, 20]. At the end of the day, as Oreskes and Belitz put it, “the most we can do is to say that a model is close to the state of the art (if it is), that it has been grounded in our best understanding of known natural processes (if it has), and that we built it on the basis of abundant, well-constrained empirical input (if we did)” [28]. The modes of a mature public encounter following from such an admission – what I address under the language of ‘virtual accountability’ – remain very much to be worked out.

3. MODELS AND WATER MANAGEMENT: A BRIEF HISTORY

Computer simulation models made their first direct appearance on the California waterscape beginning in the 1960s, appearing more or less simultaneously within the planning and operations divisions of the California Department of Water Resources (DWR). On the operations side, the build-out of the managed water network, with the addition of the State Water Project to the now nearly complete Central Valley Project, led to new technical and organizational challenges around the coordination of an increasingly complex and inter-tied network. By wiring the reservoirs in real time and developing computational models and monitoring protocols around the resulting data, project operators could more safely manage the growing complexities of a multiply inter-tied system. At around the same time, DWR and Bureau of Reclamation planners began developing a series of depletion and accretion studies simulating the effects of varying hydrological conditions on streamflow and groundwater patterns in the all-important Central Valley.

By the early to mid 1980s, these fledgling efforts had grown and achieved a measure of standardization. Efforts to link facilities connecting through the State Water and Central Valley Projects were largely complete, with real-time management information now converging on the Operations Control Center in Sacramento. Initial planning studies commissioned on a one-off basis had morphed into two more general modeling frameworks: DWRSIM, used by the California Department of Water Resources to manage the State Water Project; and PROSIM, used by the Bureau of Reclamation in its Central Valley operations. Beyond the state and federal models, many water agencies, irrigation districts, municipal suppliers and academic modelers developed simulations of their own. Some of these operated in loose coordination with the state and/or federal modeling efforts; many others were designed and run in isolation.

By the early 1990s, this heteronomy of models and modelers had become a source of both technical and political instability. The resolution of discrepancies between the state and federal models absorbed increasing amounts of time and institutional resources. A series of disputes between the DWR and local water agencies resulted in acrimonious hearings before the state water board in which modelers on either side lined up to challenge the integrity and credibility of the opposing model. Such publicly-aired rivalries within the still loosely defined professional circles of modeling gave an increasingly restive set of external actors new purchase on the previously opaque details of water management and engineering; environmentalists, agriculturalists and urban users opposed to a particular policy stance adopted by the DWR or Bureau of Reclamation were in some cases able to point to other numbers and alternative conclusions reached by different sets of formally accredited and apparently equally legitimate analysts.

At the same time, under the weight of the new technical, political and institutional demands placed upon them, the now legacy models of DWRSIM and PROSIM were beginning to break down. At the level of design, the ‘spaghetti-coded’ and idiosyncratically-produced nature of this generation of models meant that few actors other than their original creators could be said to understand their operation in enough detail to recognize and correct the frequent errors and inconsistencies that emerged in

operation. This posed, among other things, a unique personnel problem: as original modelers left the DWR or Bureau for the lucrative engineering consulting industry, government agencies found themselves in the embarrassing position of being unable to understand and run their own models (short of hiring back their own former employees at private consulting rates). At the same time, the ‘hard-wired’ nature of the models (i.e. the need to specify elements and perform changes at the level of opaque, usually Fortran, code) made comparative analyses undertaken on the basis of models both awkward and time-intensive.

Significantly, even such apparently ‘technical’ problems were embedded within and substantially owed to a wider set of institutional, political, and ecological transformations. For instance, the comparison problems posed by the hard-wired nature of DWRSIM and PROSIM could be overcome, or at least accommodated, within the relatively stable management and policy regimes inherited from the 1950s and 60s. Under such circumstances, the demands placed on the model were relatively simple: maximize deliveries and give some consideration to power generation, subject only to the constraints of flood control. This situation changed drastically as ‘environmental’ claims on the system mounted, and in particular following the passage and enforcement of the National Environmental Protection Act (NEPA) and California Environmental Quality Act (CEQA), which made Environmental Impact Reviews part of the language and responsibility of water managers throughout the state. Now, suddenly, comparative analysis was mandated by law (backed by activist pressure), and the previously only dimly experienced ‘technical’ inadequacies of the models were made glaringly apparent. Relatedly, the problem of spaghetti-coding, noted above as a personnel issue, reemerged under the pressures of the environmental movement as a fundamental problem of legitimacy: what faith should public actors place in a model whose inner workings remained opaque to all but the smallest handful of the initiated, especially in cases where such models provided the primary, even the sole, evidentiary basis for public decision-making? [18]

3.1 CalSim: consensus and controversy

Faced with such pressures, the design, practice, and politics of water models in California underwent a series of important changes in the early to mid- 1990s. Most importantly, traditional rivals DWR and BR agreed to cooperate in the joint development of CalSim, a new and widely-touted “consensus model” that would replace each of their aging proprietary models. At the level of technical design, CalSim would correct many of the shortcomings noted in its predecessors: it would be soft-wired (or “data driven”) as opposed to hard-wired, lending itself more readily to the sorts of comparative and speculative studies its predecessors were ill-equipped to handle; it would follow the now-common coding principles of structured and object-oriented programming, allowing improvements in general readability and new possibilities for modular development; it would incorporate new and standardized user interface and file management procedures that would ease the flow of data in and out of the model; it would be open source and, in principle, freely downloadable from the DWR website; and it would pull all representations of data, including the model’s crucial operating rules and assumptions, from FORTRAN code to a more flexible and accessible natural language interface, keyed specifically to the conditions and practices of western water management.

Collectively, it was argued, such changes would go a long way towards repairing the technical consensus around water management fractured by the disputes and discrepancies of the eighties and early nineties. Converging on a common model, it was also suggested, would realize new efficiencies of scale by coordinating the development efforts of DWR, BR and other modelers around a single common object. As an object shared between agencies, CalSim would grow faster, more efficiently, and more reliably [8, 11, 14].

Standing next to and supporting such technical claims were a series of explicitly political appeals. The newly open architecture of CalSim, it was suggested, could contribute importantly to public confidence in the tool, establishing a form of political legitimacy that its predecessors had at first not needed, and then distinctly lacked. With this common understanding in place, it was hoped, a substantial portion of the legal and political controversies that had embroiled the system since the rise of the environmental movement could be done away with, and the various parties (but most especially the managers in the DWR and Bureau) could get back to the work of the rational management and distribution of resources. In this regard, CalSim was touted as the putative lynch-pin in a peace-through-science settlement promising to restore both order and a measure of civility on the California water system.

Underlying all of these hopes lay the common techno-political dream of *transparency*. Through such innovations as structured programming, natural language interfaces, standardized file management procedures, and modular (object-oriented) development strategies, CalSim aspired to a level of architectural transparency far surpassing the opaque code of its predecessors. Such innovations, it was argued, would improve model reliability, supportability, and cross-agency technical collaboration. But if transparency was an architectural virtue, it was also a democratic one: in the fractious climate of California water politics in the early 1990s, the very ‘openness’ of the model – open source, open code, open documentation – was heralded as an important and necessary *political* accomplishment. On both the architectural and democratic fronts, however, transparency was a partial and tenuous achievement at best, consistently undermined by the challenges of data and organizational alignment and the in some ways irreducible complexity of the system under representation. More subtly, the principle of transparency, while commonly presented as a solution to the absence of trusted relations, turned out to depend on them. From this perspective, the ‘external’ transparency of models rested substantially on the ‘internal’ stabilization of its constituent parts through the principled agreement to leave certain assumptions and assertions (including the professional competence and good faith of its practitioners) unquestioned. Absent this level of stabilization, as subsequent events would show, the promise of transparency failed and models could be rendered once again vulnerable to critique, dispute, and the real-world machinations of western water politics [18].

4.0 THE TROUBLE WITH NUMBERS

4.1 The State Water Project Reliability Report

In August of 2002, the California Department of Water Resources released a draft document with the apparently innocuous title of “The State Water Project Delivery Reliability Report.” The task of the report was seemingly straightforward: to “provide current

information on the ability of the SWP to deliver water under existing and future levels of development, assuming historical patterns of precipitation” [3]. This exercise in predictive modeling, undertaken using CalSim and the 73 years of annualized data contained within the acknowledged period of record, was complicated by two contextual factors. First, the report itself grew out of (and was indeed mandated by) ongoing legal controversies surrounding the 1995 Monterey Amendments to the State Water Project contracts, which environmental groups charged with fundamentally over-stating the delivery capacities of the state water system – the aptly-named problem of ‘paper water’ – thereby throwing open the door to unrestricted and unsustainable growth in the state. Second, recently passed bills in the California Assembly requiring private land developers and local planners to demonstrate water supply reliability twenty years into the future had essentially granted CalSim – the only tool deemed capable of making this sort of prediction – the weight of law. Within this politically-charged climate, reactions to the DWR’s projections were swift. In written and verbal testimony submitted as part of the Report’s public review and comment period, the modeling on which the analysis was based was attacked as deficient on a number of grounds: for failing to account for the potentially serious effects of climate change on regional water supplies; for its inadequate attention to water rights senior to the State Water Project (including native, municipal, count-of-origin, and public trust claims) which could limit future deliveries through the state system; its insufficient representation of dynamics within both the federal portion of the system and the conjoined groundwater system; and for its decision to hold regulatory constraints on the system constant, thus failing to account for the likelihood of either future infrastructural development that would increase project supply capacity, or future endangered species claims that would effectively reduce it.

Arguably, none of these detailed and somewhat arcane technical debates would have entered the public sphere at all, were it not for two of the report’s central, and deeply counter-intuitive, findings: first, that delivery reliability would actually *improve* over the course of the 2001-2021 period, in spite of the increasing upstream demands placed on the system; and second, that the SWP could be relied upon (at both 2001 and 2021 levels of development) to deliver water at levels that were, on average, nearly *fifty percent higher* than historic deliveries. In contrast to real-world SWP deliveries hovering in the neighborhood of slightly more than 2.0 million acre-feet (maf) per year, the report announced simulated deliveries ranging, on average, from 2.96 to 3.13 maf (at 2001 and 2021 levels of development, respectively). By what logic, asked the report’s many critics, might the constraints on an already overtaxed and still tightening system be expected to *ease* over the next twenty years, at a moment when virtually everyone in the California water community was predicting and preparing for a much darker scenario of growth, shortage, and conflict? As one critic noted, somewhat incredulously, “We are asked to believe that the SWP will reliably, on average, provide an additional million acre feet of water (50% greater than past performance). The finding defies logic and is inconsistent with the system’s actual performance.” [35].

These questions took on added political weight when Senator Michael Machado, member of the powerful Senate Committee on Agriculture and Water Resources, wrote to express his concerns.

Noting the widespread public backlash and tension surrounding the draft document and the modeling underlying it, Machado argued that the report was ‘premature’ and urged DWR to take active steps to address the concerns and criticisms leveled against it. As noted by Machado, the stakes went well beyond the report itself:

- Local development could be hampered if, when complying with SB 221 (Kuehl) and SB 610 (Costa) of California’s Environmental Quality Act (CEQA), there are significant disputes over current and future water supplies.
- Conclusions of CalFed’s Integrated Storage Investigations (ISI) will be suspect given that the same model is used in both the ISI and the SWP Reliability reports.
- Future statewide bonds for increasing water supply will be in jeopardy, if opponents can credibly challenge the underlying analysis.

As a measure of his concern, Machado took the unusual step of asking the California Research Bureau, the research arm of the state library system, to produce a formal analysis and comment on the report. The CRB commentary, drafted by Assistant Director Dennis O’Connor, constituted the longest and most detailed intervention over the course of the SWP reliability report controversy, weighing – and in part endorsing – the claims of the report’s environmental critics² [26].

Dominating the public hearings and comment period, as feared by Machado and detailed in the CRB report, were public concerns and criticisms around CalSim itself. These were organized around two interlinked questions: first, the manifest (but underacknowledged) limitations of the model in use, as shown up in the specific context of the report’s production; and second, the credibility of models as policy tools more generally, particularly where supplying the primary evidentiary input to public decision-making. On the first point, critics were quick to point out the remarkable poverty of the model when it came to representing important facets of the California water system – most notably, surface-groundwater interactions and delivery curtailments following unacceptable takes of endangered species at project pumping facilities – that must play a significant role in all future projections of real-world supply availabilities. Similarly, critics noted, CalSim assumed a level of order and predictability in the *human* operation of the system that was not always in evidence; complex operational rules and legal infrastructure aside, the various actors in the system, from DWR operators to project contractors, did not always follow the logics and rationalities that the model (though in principle, also their operating and

² Specifically, the CRB report questions the Department’s suggestion that the reason historical deliveries fall well below modeled results is that the contractors haven’t requested their full allotments in past; the apparent failure to account for the effects of upstream development (and thus consumptive use) under the 2021 scenarios; the weak representation of key variables such as groundwater interactions; the discrepancy between CalSim monthly results and the more modest projections associated with finer-grained daily models; and the use of the model for predictive rather than comparative purposes.

contractual obligations) imposed upon them. This was particularly true when the stakes were highest (most notably, under drought conditions) when operators and other actors were required to cobble together local responses to crisis that strayed from and sometimes violated the formal operating procedures laid down in the model. For all these reasons, CalSim was too simple, too narrowly framed, and entirely too thin to incorporate the degree of nuance and complexity needed to reliably project the real-world futures of the system.

Beyond this, critics felt that the DWR was essentially asking them to accept *on faith* the efficacy of a rather opaque technological artifact to which they had been granted little effective access, let alone participation. Critique returned time and again to the fact that at the time of the report, CalSim had yet to undergo anything like the sort of rigorous external assessment that might establish some grounds for its legitimacy, despite the fact that it had then been in use for several years by analysts within the department, the Bureau of Reclamation, and a number of informally affiliated private consulting firms. Critic after critic noted that CalSim had yet to be tested against the historical record in anything like the sort of validation or calibration exercise typically expected of models in other domains of science and public life. Without knowing that the model could effectively reproduce the intensely-observed history of water in California, why should those suspicious of its claims (and the motivation of those making them) grant their assent? Similarly, critics noted, CalSim had never been reviewed in anything like a systematic way by anyone other than its creators and principle users within the department, bureau and a few hand-picked consultants. Given its centrality to water planning and management in the state, why hadn't a formal peer review been conducted? In the absence of these initiatives and/or the presumption of good faith on the part of the department, there appeared to be little left to compel general assent to the model and its claims. In this context, as one critic noted sardonically, "'Our model says so' is not enough to base policy on" [35].

By the end of the SWP reliability report public comment period, the concern cited in Senator Machado's letter (and privately conceded by modelers and planners in the Department of Water Resources) had been borne out: passing more or less quickly over the details of the report, public controversy had come to rest squarely on the credibility of CalSim itself. In the process, previously arcane details of model design and operation were discussed, debated, and sometimes challenged by actors well beyond the usual core of modelers, engineers and departmental managers. Such events contributed to a larger mood of public skepticism that disrupted and arguably put to rest the hopes associated with the peace-through-science settlement of the 1990s.

4.2 The 2005 Water Plan Update

While concerns around CalSim and the credibility of models more generally were surfacing in the context of the SWP Reliability Report, models were being opened up from a different direction in the context of statewide water planning. The immediate venue for this debate was the California Water Plan, a once every five year attempt to square the circle of California water policy, synthesizing the starkly different interests of agricultural, urban, environmental, and other policy claimants into a credible and workable statewide management framework.

The most recent water plan update process, which issued its draft report for public comment in summer 2005 (a full two years behind its legally mandated schedule) grew out of a particularly acrimonious set of debates culminating in the bitterly divisive water plans of 1993 and 1998. As the record of public comment reveals, commentators on the earlier plans differed sharply as to the nature of the crisis unfolding in California. For a range of environmental critics, planning was urgently required to redress the great ecological disaster long unfolding in the Bay-Delta and other areas of the state. For agricultural supporters and some urban water agencies, the crisis was precisely reversed, namely, that the accumulated weight of population growth and environmental demand had rendered the water supply system fundamentally unreliable and in particular vulnerable to future fluctuations of the hydrological cycle. Not surprisingly, respondents also varied in their estimation as to where the elusive 'new water' needed to fix the California system would come from: urban and agricultural contractors in the state looked for the most part to new surface storage and conveyance facilities, while environmental advocates argued in favor of so-called 'soft path' strategies, in the form of efficiency gains, strict conservation measures, and other sorts of demand-side management strategies [36]. Complicating all these debates were the looming effects of regional climate change, feared by many to reduce the natural system's capacity to carry over winter precipitation into the peak summer months of agricultural, urban, and power consumption. [21].

These varied policy positionings, which cut to the heart of the future growth and development strategies of the state, returned time and again to apparently technical disputes around the description, quantification, and prediction of water. Scientific understanding of groundwater depletion and recharge processes was appallingly bad, charged some commentators, leading to overdraft figures, projections, and overall water balances that were little better than guesses. The plans substantially overestimated future agricultural demands, charged others, underplaying the effects of future efficiency gains and the water-saving potential of continuing market-driven 'ag-to-urban' transfers. The plans similarly exaggerated future urban uses, others argued, which were based on demographic projections that failed to consider the potential dampening variables of economic recession, land price inflation, etc. In both cases, the failure to assign real-cost pricing – i.e. modulating demand projections according to the rising prices that would (or should, in the absence of ongoing subsidies to agriculture) accompany future water scarcities – significantly skewed both urban and especially agricultural demand in an upward direction. Estimates of present and future 'environmental water' needs – counted as a line item for the first time in the 1993 update – were argued to be either too high or too low, and in any case inadequately specified and/or based on a level of scientific understanding insufficient to justify the large-scale restructuring of project deliveries [18].

It was against this acrimonious backdrop that work on the current water plan began. Vowing to redress the participatory failings of 1998, DWR retained professional mediation services and expanded the plan's official Advisory Committee to 65 people, including for the first time district level, tribal, and environmental justice representatives. At the level of content, three consequential decisions were taken. First, in an effort to acknowledge the deep uncertainties facing water prediction and management in the state,

the controversial single figure “gap analysis” of past plans (subtracting current supplies from projected needs, and proposing facility or management changes to redress the balance) was dropped in favor of a ‘scenarios’ approach, in which the performance of the system under multiple future supply and demand conditions would be contemplated. Second, in the face of widespread skepticism surrounding the Department’s procedures for normalizing data into water year ‘types’, the advisory committee elected to work from real data sourced from three recent water years: 1998 (classed as a wet year), 2000 (“average”), and 2001 (the driest on record since the 1987-1992 drought). Arguably the most significant development, however, came with the decision to abandon numbers altogether – or rather, to put off the thorough processing of them until later stages of the plan. In a sharp departure from prior plans, Bulletin 160-03 would be issued sequentially, with a policy-focused first phase describing the current system state and describing general priorities and potential policy stances, a second tool-building phase establishing in more detail the precise approaches to be taken in quantifying the system, and a third phase in which the qualitative scenarios outlined in phase one would be populated with data and at last calculated out. There were some immediately pragmatic reasons for adopting this strategy. By early 2003, the plan was far off its timeline and showed little or no hope of hitting its scheduled release date at the end of the year; the turnover of senior personnel within the Department following the gubernatorial recall election of 2003 had recently introduced additional delays and uncertainties. There was also some sense, shared among DWR officials and advisory committee members involved in the planning process, that putting off the zero-sum game of calculation, like the scenario decision before it, may have softened the sharper edges of interest group conflict and therefore played a role in keeping stakeholders committed and engaged in the planning process.

At the most basic level, however, the decision to prepare and release the plan in stages was owed to widespread reservations around the quality and trustworthiness of numbers. Through the early stages of planning, the broad lines of quantitative disagreement, like the political split more generally, followed those laid down in the aftermath of the 1998 plan. Environmental groups contested DWR procedures for calculating urban and agricultural demand (in particular, its utter neglect of price signals), pushed for new ways of calculating the savings to be achieved through urban and agricultural efficiency, and urged the state to adopt beneficiary-pays and true-cost pricing principles. Agricultural groups and urban suppliers argued that the numbers on “new water” produced through urban and agricultural efficiency improvements were wildly optimistic, and suggested that many of the ‘soft’ gains to be had by such measures had already been achieved during the 1987-1992 drought (the so-called ‘demand hardening’ argument). Agricultural representatives argued further that such projections implicitly endorsed an expanded and ultimately short-sighted program of ag-to-urban water transfers that would leave the state unable to meet its own ‘food and fiber’ needs within the foreseeable future. By 2003, as the plan’s official delivery date neared, numeric and political tensions on the advisory committee deepened. Following early efforts to avoid the traditional sectional splits into farm, city, and environment – as one respondent later noted, “we were trying at the start to not go positional” – around the middle of 2003 the

committee regrouped itself into caucuses, with representatives now speaking on behalf of the traditionally-identified groups.

Through this process of asserting and disputing the adequacy or otherwise of specific numbers, the advisory committee gradually came to a more general awareness of the limits and problems of data and models in general. In 2002, following widespread expressions of concern within the advisory committee over the credibility of the models on which the 2030 projections were to be based, several members of the advisory committee formed a Modeling Work Group, dedicated to the task of exploring and reporting back to the group on the strengths and limitations of available data and modeling frameworks. In September 2002, the group prepared a formal modeling proposal which was subsequently adopted by the advisory committee. In contrast to the technocratic certainty characterizing the language of previous water plans, the advisory committee statement struck a pointedly skeptical note. While acknowledging that the “proposed models have some constructive role to play in Update 2003,” the work group cautioned that “the potential exists for policy makers and the legislature to misuse modeling data, which necessitates judgment in releasing select results and identifying model limitations” [5]. Beyond this,

Models are inherently uncertain. Any decisions based on models should include this caveat. All models in Update 2003 have limits: DWR staff and the Advisory Committee will identify those limits in the main plan’s text and provide details in the appendix. The Advisory Committee will bear such limits in mind and reflect on improvements when interpreting the results of model runs. [5]

Far from a ringing endorsement or a blanket condemnation, the response of the advisory committee to the presentations of DWR modelers was both critical and pragmatic. The members of the committee (and in particular its modeling work group) were willing to acknowledge the usefulness of models as a potential input to policy-making, but were not willing to grant their assent on faith, or to cede to model results the preponderant weight in future water decisions. In the face of such ongoing uncertainty, advisory committee members were urged to rely on their “collective wisdom,” and treat the predictive claims of the models with a degree of informed skepticism.

Such skepticism also became the occasion for a fundamental rethinking and the beginnings of a redesign of the state’s modeling infrastructure. As members of the workgroup noted, the existing suite of models and numeric analysis tools was significantly, perhaps even dangerously, misaligned with the sorts of questions water managers and public decision-makers in the state were increasingly being called upon to address. As one advisory committee member noted in a letter to the committee,

Most planning analysis and data collection for California’s statewide water resources were developed for an era of large-scale water facility development. Our analysis capability continues to specialize in the operation and planning of the large State and Federal water projects. Most analysis largely neglects the local and regional activities which are the

hallmark of current water management, such as water conservation, conjunctive use of ground and surface waters, water transfers, and wastewater reuse. [25]

On this basis, “DWR’s data collection and analysis capabilities must be substantially re-directed and re-engineered to re-orient DWR planning to aid, support, and integrate local and regional efforts.” [25]

In the end, the perceived weakness of the available numbers and models led the advisory committee and DWR planners to the three-phase approach noted above: they would produce a plan, but it would, at least in the interim, contain very few numbers, and certainly none of the summative sorts of numbers associated with things like the reviled ‘gap analysis’ of past plans. At the same time, work would begin on a second phase, in which current data and model deficiencies would be identified, and long-range approaches to correcting these undertaken. Armed with the new numbers and tools, the plan’s third phase would at last ‘cost out’ numerically the scenarios generated qualitatively in phase one. The draft of the long-awaited plan’s first phase was released for public comment during the summer of 2005 (where, predictably enough, its lack of numbers came up for regular criticism). The third and final phase is now projected to arrive in 2008 – five years late, and in precisely the year the *next* water plan was to have been delivered.

5.0 VIRTUAL ACCOUNTABILITY

The efforts of the water plan advisory committee and its modeling workgroup represent only one part of a larger effort to address and restore the credibility of models and modeling as input to public policy that has been arguably damaged by recent controversies in California water management. Stung by criticisms received during the Monterey Amendment / SWP Reliability Report controversies, the DWR, acting in conjunction with the Bureau of Reclamation, has taken several steps to address at least the most immediate concerns of its critics. In 2003, the DWR issued the findings of its Historical Operations Study, the most serious attempt to validate CalSim vis-à-vis the historical record to date. Running the model against operating logics, streamflow data, and water delivery records from 1975-1988, the study authors argued that CalSim in fact performed remarkably well, returning Delta outflow figures that differed on average by only 7% from historical values, and hitting within 5% during the crucial drought period of 1987-1992. On the question of groundwater, where hard data was (and remains) notably lacking, CalSim was tested against the more detailed representation contained within the Central Valley Groundwater Surface Water Model and found to be broadly, though not perfectly, compatible [4]. In March 2005, DWR officials presented preliminary results from the first large-scale CalSim sensitivity analysis, in which the model again performed reasonably, though not perfectly, well.

In arguably the most significant development to date, in 2003, the CALFED Science Program (responding to DWR requests) undertook the first large-scale (if still limited) peer review of CalSim, conducted by seven ‘external’ experts chosen for their long experience in water operations and simulation modeling. The results of the review were mixed: while endorsing the overall technical soundness and general approach of the model (praising in particular CalSim’s open source and consensus-building

ambitions), a wide range of reservations were expressed: geographic coverage in the model was weak and notably incomplete, in particular with regard to Southern California and Colorado River transfers; attention to questions beyond traditional supply concerns was weak or non-existent; the distributed character of model development within and beyond the agencies raised important questions of versioning, consistency, and quality control; the model’s understanding of real-world operational dynamics and decision-making was grossly and unacceptably simplified; and despite apparently real and laudable aspirations towards openness, insufficient effort (in the form of documentation, user-friendly interfaces, user support, public workshops, etc.) had been devoted to making the model usable or even comprehensible beyond the confines of a fairly narrow circle of experts [8, 14].

As this list of design responses to controversy (and their shortcomings) may begin to suggest, the real-world challenges of ‘modeling democratically’ within realms of complex and bitterly contested public policy are immense. Moreover, they spill regularly and confusingly beyond the confines of ‘straight’ technical practice into broadly sociological registers of trust, confidence, and credibility which modelers and water managers are ill-equipped by training to deal with (though many have gained considerable practical skill in this regard). Under such circumstances, technical ‘fixes’ to political problems are likely to fall short of their goals (as indeed are ‘sociological’ responses to hard technical concerns). The challenge, as always, is to work across the two sides of this divide simultaneously.

What might an appropriately deliberative solution under such circumstances – what I’m describing under the language of ‘virtual accountability’ – entail? First, it should be noted that other modeling frameworks, some more supportive of stakeholder deliberation, may be identified. ‘Gaming’ models have been developed and successfully deployed in several instances as an aid and heuristic to contentious group decision-making processes. ‘Screening’ models (including periodic calls for a ‘CalSim-lite’) may be developed and deployed in forms that sacrifice a degree of analytic precision and granularity, but may gain in broader stakeholder accessibility and general analytic wieldiness. Neither of these approaches could entirely supplant the multiple functions CalSim is currently called upon to perform; but they could perform at least some of those functions in a more deliberatively-supportive and ultimately effective manner.

Second, as noted by respondents to the CalSim peer review, modelers could also do more to build effective public access to their tools and findings, through better documentation, interface development, public training and information sessions, and other sorts of mechanisms. Some of these are already underway, e.g. regional review workshops fielding public comment around the representation of particular system components within the larger CalSim architecture. This sort of public investment, in what we might think of as ‘translation goods,’ ought to be undertaken as democratic as well as purely analytic investments – a point that funding realities and structures within the DWR are currently ill-equipped to accommodate.³

³ Notably, funding for CalSim is organized under the ‘project’ side of the DWR organizational hierarchy, concerned primarily with the planning and operation of the State Water Project and

Third, as painful and inefficient as it may sometimes be to move beyond the measured world of technical decision-making, serious and sustained efforts at broad public engagement seem the most promising road to the longer-term goals of widespread model literacy and trust that in the end will be needed to sustain and extend the viability of modeling as a policy technology. Important early movements in this direction may be identified in the work of the water plan advisory committee. Until California's water problems go away and/or simulation modeling achieves an unquestioned sophistication and place within the pantheon of credible policy knowledges – neither of which seems likely to happen anytime soon – ongoing efforts to engage across the technical-public divide remain the most likely long-term strategy for building trust, confidence, and legitimacy – and ultimately an effective and democratically sensible water policy.

Beyond their immediate implications for simulation modeling in the water policy context, the story sketched above speaks to issues of wide and growing digital government concern. While the entrenched conflicts, high stakes, and deep uncertainties dominating California water policy may make this an unusually intense laboratory for the observation of simulation in action, it is by other measures an old and entirely unremarkable story: actors on all sides of the California water debate are by now well-versed in the challenges of performing technical work in a politically fraught field – the ever present challenge of what I've described elsewhere as ‘doing hard politics with soft numbers’ [18]. This common tension and dynamic has yet to receive the theoretical elaboration it deserves, in either the digital government or more general public policy literatures.

There are also evolutionary dynamics worthy of digital government and broader public policy note. In each of the controversies sketched above, the apparently technical art of modeling was opened, however awkwardly and painfully, to review and criticism beyond its immediate circle of expertise. In the process, public light was cast into areas of practice formerly ceded almost entirely on trust to a domain of professional expertise. This occurred not through any abstract notion of participation or transparency (though it has clearly and regularly drawn on such resources), but rather through the hard and contingent work demanded in complex and contentious political settings. In this regard, the exigencies of the political field brought out, exploited, and in some measure created latent instabilities in the technical field. But once opened, such controversies were not easily or quickly resolved, precisely because of their tendency to spill across the conjoined worlds of technical and political action.

Second, despite the arguably distinctive intensity of the California water case, the general presence of simulation tools at the center of contested domains of public policy seems unlikely to decrease in future (though given the challenges and instabilities noted above, this is not a foregone conclusion). The rapid and comparatively recent development of simulation techniques and their generally speedy diffusion through the policy field to date would seem to suggest that ‘model knowledges’ – conclusions, predictions, and other assertions of fact drawn partly or primarily

only secondarily with the wider concerns around water policy and participation traditionally housed in the Department's planning division.

on the basis of computer simulations – are likely to figure as more rather than less significant policy technologies in future. Under such circumstances, the sorts of model literacy advocated above may become an increasingly important attribute and skill-set, both within digital government scholarship and democratic polities more generally.

Third, as the case study suggests, the work required to build and sustain models as meaningful objects in the world comes in many forms: conceptual, mathematical, and computer-based; but also organizational, political, and broadly sociological. The latter is sometimes treated as an add-on to the real work of modeling, perhaps necessary but fundamentally distinct from the technical work of building and running the ‘models themselves’ (a form of what we might call ‘code realism’). In the world of California water modeling, however, such distinctions are hard and arguably becoming harder to maintain. As the field study traced above will begin to suggest, it turns out to be extraordinarily difficult to assign where, precisely, the work of modeling begins and ends. Can a model be reduced to code? To data? To the immediate network of designers and decision-makers that build and use it? To the wider networks of trust and credibility that sink or sustain its claims? The diversity of ecologies in which models build and hold meaning give them, like other many other complex artifacts, a considerable degree of ‘ontological sprawl’ that is neither easily nor obviously reduced. Effective modeling, like other instances of technical work in the policy arena, must attend to this diversity by meeting head-on the full range of technical, institutional, and broadly sociological conditions that enable and constrain its work.

6. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 0131923, and by NSF Dissertation Improvement Grant No. 0425261. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation. The author wishes to thank Geoffrey Bowker, Robert Horwitz, Jane Fountain, David Lazer, and Sheila Jasanoff for guidance and support during the research and writing of this project.

7. REFERENCES

- [1] Beven, K. Changing Ideas in Hydrology: The Case of Physically-Based Models. *Journal of Hydrology* 105 (1989), 157-172.
- [2] Beven, K. Prophecy, reality and uncertainty in distributed hydrological modeling. *Advances in Water Resources* 16 (1993), 41-51.
- [3] California Department of Water Resources. *The State Water Project Delivery Reliability Report*. Sacramento, Aug 2002.
- [4] California Department of Water Resources. *CalSim II Simulation of Historical SWP-CVP Operations: Technical Memorandum Report*. Sacramento, Aug 2003.
- [5] California Water Plan Advisory Committee Modeling Work Group, Update 2003 Modeling Proposal Addendum (email to Advisory Committee). Accessed Aug 9 2005 at: <http://www.waterplan.water.ca.gov>

- [6] Canham, C., Cole, J., and Lauenroth, W. Models in Ecosystem Science. In *Models in Ecosystem Science*, eds. Canham, C., Cole, J., and Lauenroth, W. Princeton University Press, Princeton, 2003, 1-12.
- [7] Carpenter, S. The Need for Fast-and-Frugal Models. In *Models in Ecosystem Science*, eds. Canham, C., Cole, J., and Lauenroth, W. Princeton University Press, Princeton, 2003, 455-460.
- [8] Close, A., Haneman, W., Labadie, J., Loucks, D., Lund, J. McKinney, D., Stedinger, J. A Strategic Reivew of CALSIM II and its Use for Water Planning, Management, and Operations in Central California. Panel report submitted to the California Bay Delta Authority Science Program Dec 4, 2003.
- [9] Collins, H. *Changing Order: Replication and Induction in Scientific Practice*. University of Chicago Press, Chicago, 1985.
- [10] Cottingham, K. et. al. Increasing Modeling Savvy: Strategies to Advance Quantitative Modeling Skills for Professionals within Ecology. In *Models in Ecosystem Science*, eds. Canham, C., Cole, J., and Lauenroth, W. Princeton University Press, Princeton, 2003, 428-436.
- [11] Draper, A., Munevar, A., Arora, S., Reyes, E., Parker, N., Chung, F., Peterson, L. CalSim: A Generalized Model for Reservoir System Analysis. Unpublished paper available at: <http://science.calwater.ca.gov/index.html>.
- [12] Edwards, P. Global Climate Science, Uncertainty, and Politics: Data-Laden Models, Model-Filtered Data. *Science as Culture* 8:4 (1999), 437-472.
- [13] Ezrahi, Y. *The Descent of Icarus: Science and the Transformation of Contemporary Democracy*. Harvard University Press, Cambridge, MA, 1990.
- [14] Ferreira, I., Tanaka, S., Hollinshead, S., and Lund, J. *CALSIM II in California's Water Community: Musing on a Model*. Report prepared for the CALFED Science Program, Jan 20, 2004.
- [15] Fox-Keller, E. Models of and models for: theory and practice in contemporary biology. *Philosophy of Science* 67 (2002), S72-82.
- [16] Galison, P. *Image and Logic: A Material Culture of Microphysics*. University of Chicago Press, Chicago, 1997.
- [17] Gusterson, H. *People of the Bomb: Portraits of America's Nuclear Complex*. University of Minnesota Press, Minneapolis, 2004.
- [18] Jackson, S. *Building the Virtual River: Numbers, Models, and the Politics of Water*. Unpublished Ph.D. Dissertation, University of California, San Diego, 2005.
- [19] Jasianoff, S. Technologies of Humility: Citizen Participation in Governing Science. *Minerva* 41 (2003), 223-244.
- [20] Jasianoff, S. *Designs on Nature: Science and Democracy in Europe and the United States*. Princeton University Press, Princeton, 2005.
- [21] Kiparsky, M., and Gleick, P. Climate Change and California Water Resources: A Survey and Summary of the Literature. Report by Pacific Institute for Studies in Development, Environment, and Security, July 2003.
- [22] Kraemer, K., Dickhoven, S., Tierney, S. and King J. *Datawars: The Politics of Modeling in Federal Policymaking*. Columbia University Press, New York, 1987.
- [23] Kwa, C. Modeling Technologies of Control. *Science as Culture* 4:20 (1994), 363-391.
- [24] Lauenroth, W., Burke, I., and Berry, J. The Status of Dynamic Quantitative Modeling in Ecology. In *Models in Ecosystem Science*, eds. Canham, C., Cole, J., and Lauenroth, W. Princeton University Press, Princeton, 2003, 32-48.
- [25] Lund, J. In the dark, all cats are gray (letter to the Department of Water Resources, Division of Planning and Local Assistance, Sep. 6, 2003).
- [26] O'Connor, D. Comments on the Department of Water Resources' Draft State Water Project Supply Reliability Report, Nov 1, 2002.
- [27] Oreskes, N., Shrader-Frechette, K., and Belitz, K. Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences. *Science* 263:5147 (Feb. 4, 1994): 641-46.
- [28] Oreskes, N., and Belitz, K. Philosophical Issues in Model Assessment. In *Model Validation: Perspectives in Hydrological Science*, eds. Anderson, M. and Bates, P. John Wiley and Sons, New York, 2001, 23-42.
- [29] Porter, T. *Trust In Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton University Press, Princeton, 1995.
- [30] Shackley, S. and Wynne, B. Global Climate Change: The Mutual Construction of an Emergent Science-Policy Domain. *Science and Public Policy* 22:4 (1995), 218-230.
- [31] Shackley, S. and Wynne, B. Representing Uncertainty in Global Climate Change Science and Policy: Boundary-Ordering Devices and Authority. *Science, Technology, and Human Values* 21:3 (1996), 275-302.
- [32] Sismondo, S. Models, Simulations, and Their Objects. *Science in Context* 12 (1999), 247-260.
- [33] Star, S.L., Grisemer, J. Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science* 19 (1989), 387-420.
- [34] Thevenot, L., Boltanski, L. *On Justification: Economies of Worth* (tr. Catherine Porter). Princeton University Press, Princeton, 2006.
- [35] Wilkinson, R. Comments on DWR's Draft Report 'The State Water Project Delivery Reliability Report', Oct 31, 2002.
- [36] Wolff, G., and Gleick, P. The Soft Path for Water. *The World's Water 2002-2003: The Biennial Report on Freshwater Resources*. Island Press, Washington, 2002, 1-32.

Organic Development: A Top-Down and Bottom-Up Approach to Design of Public Sector Information Systems

Michael Tyworth

The Pennsylvania State University
311B IST Building
University Park, PA 16802
01-814-571-0585
mtyworth@ist.psu.edu

Steve Sawyer

The Pennsylvania State University
301F IST Building
University Park, PA 16802
01-814-865-4450
sawyer@ist.psu.edu

ABSTRACT

In this paper we lay out interim findings and speculate on the implications for practice and theory of integrated criminal justice systems in law enforcement. In doing this we theorize on public sector information systems and their uses of information and communication technologies as engaging in what we call “organic development.” To develop our theorizing on organic development, we draw on a field study of the San Diego, California area’s Automated Regional Justice Information System (ARJIS). We develop organic development as drawing on both top-down and bottom up approaches to engaging the technologies, technological infrastructures, governance principles, and work practices that, together, are an integrated system.

Keywords

Institutional Theory, Integrated Criminal Justice Systems, Social Informatics, Emergent Design, Organic Design

1. INTRODUCTION

What technology architectures, governance structures and work practices are associated with successful integration of information and communications technologies (ICT) into law enforcement? In this paper we lay out interim findings and speculate on their implications relative to this question, pursuing more coherent theories, and greater insights into the practices, of integrating ICT into public safety, criminal justice and law enforcement. And, we theorize on public sector information systems as engaging in what we call “organic development.”

Law enforcement agencies have long recognized the need to integrate their ICT both within and across their individual organizational boundaries [12, 21]. Having recognized the need to integrate, law enforcement agencies and sponsors continue to struggle translate these needs into an operational reality [13-15, 25]. More recently, ICT integration became a top priority for homeland defense policymakers and renewed emphasis has been given to developing new technologies and forms of organizational collaboration [38]

The result has been a seeming explosion of design and development initiatives nationally, many with their own unique, and in some cases competing, approach to addressing the problem of organizational and system integration. These initiatives have been termed integrated criminal justice information systems

(ICJS). Examples of such initiatives include the Capital Integrated Wireless Network (CAPWIN) undertaken in the Washington D.C. metro area; Pennsylvania’s Justice Network system (JNET); Charlotte-Mecklenburg’s Knowledge-Based Cops (KBCOPS) system; and the Automated Regional Justice Information System (ARJIS) being developed for use in the San Diego, California region. There are many others.

The premise driving our work is that learning and sharing from these initiatives will lead to improved integrated criminal justice systems. To do this demands studying – to understand -- the proper governance structures, successful work practices and appropriate technological infrastructures. If the policy goal is indeed integration of law enforcement ICTs on a national scale, then it is critical to identify those designs that are successful and those that are not in order to avoid simply the development of new generation of systems that are as “siloed” as the systems they are intended to replace.

In this paper we present preliminary findings from our ongoing case study of the Automated Regional Justice Information System (ARJIS). We focus in particular on ARJIS’ “organic” or emergent method of system design and development as one that has lead to successful system outcomes. This is in direct contrast to other ICJS design and development practices that emphasize a “grand” or enterprise approach to design where the system is designed in its entirety prior to development. The remainder of this paper consists of a review of the relevant literature, a description of our ongoing study of ARJIS and a discussion of this emergent design approach.

2. LITERATURE REVIEW

We begin by noting the entrenched but chaotic roles of ICT in policing; a historical practice of uncoordinated development efforts; and the need for integration of systems across organizational boundaries. We review our intellectual frame, Social Informatics: a perspective that focuses on our attention to the socio-technical and contextualized nature regarding the design, uses, and consequences of ICTs [32]. We then highlight our theoretical approach as grounded in institutional theory: a view that engages institutions as comprised of regulative, normative, and cognitive structures that provide and define social meaning [18].

2.1 Policing & Technology

For over 70 years, law enforcement agencies have engaged ICT for crime analysis, crime prevention, and agency administration. An early use of ICT in law enforcement was the combination of a simple map covered with push-pins to denote crime activity [29]. From maps and pins; the law enforcement community has gone on to adopt a variety of different ICT including radio, cellular phones, wireless computing, computer-aided dispatch (CAD), and electronic records management systems (RMS) [12]. Using ICT in law enforcement is approaching ubiquity and is now considered to be an fundamental component of policing [17].

These ICTs are used in many, legitimate, ways. For example, an officer patrolling a beat may have the dispatcher run a license plate through the RMS to check if a vehicle is stolen. A detective may run a license plate number through an RMS to see if the vehicle has been used in other crimes; and a crime analyst may mine the data in an RMS looking for patterns of criminal activity [27].

The increased reliance on ICTs in law enforcement raises a number of issues. Perhaps the most pressing issue resulting from the incorporation of ICTs into policing is that system design and development has for the most part been done in an ad hoc and incompatible manner [12, 26]. This piecemeal approach to system design and development has resulted in law enforcement agencies being burdened with inflexible-but-entrenched systems that are generally incompatible with other law enforcement systems outside of the organizational boundaries, and occasionally even incompatible with other systems within the organization itself! Incompatibility among ICTs in law enforcement not only impacts tactical operations by hampering the sharing of mission critical information; but it also hampers information-sharing initiatives such as the National Crime Information Center (NCIC).

Another factor contributing to the lack of systems integration across agencies is the nature of policing in the United States. Law enforcement in the U.S. is structured around a federalist model both in terms of the relationship between the federal government and the states, and within the states themselves. In a federalist model each level of government has its own jurisdiction. At the federal level there is the Federal Bureau of Investigation (FBI), the Border Patrol, U.S. Attorney Office, U.S. Marshals, U.S. Secret Service, and the police agencies of the various military services. The jurisdictions of these agencies are interstate and international and the focus is on the prevention and prosecution of federal and federal-level (interstate/ international) crimes. At the state level there is the state police, and the occasional state marshal service (e.g., the Texas Rangers). The predominant role of state police is traffic law enforcement on the state highways; however state police also enforce other state level laws such as customs and organized crime [30]. At the local level there can be county, city, township, and even institutional (e.g., university) law enforcement agencies all within a limited geographical region. These agencies are engaged in the enforcement of local

laws (including traffic), as well as prevention and solving of local crimes. What this means is that in the United States, the primary for delineating governmental and organizational boundaries is geographical location.

Take for example, the relatively small town of State College, Pennsylvania. State College is a university town with a population ranging from about 40,000 during the summer to 80,000 during the academic year [2, 3]. Agencies that have jurisdiction in the State College area five township police departments, the borough of State College police department, the Pennsylvania State University police department, the Pennsylvania State Police, the Pennsylvania Domestic Relations Service, Corrections, and Game Wardens, and the Centre County Sheriffs department [1]. This totals to 12 different police agencies in a relatively small geographical area, each with its own jurisdiction, management structure, funding structure, organizational goals, and ICT infrastructures. Moreover, the police forces of the four adjoining townships are cross-sworn with State College Police (as are the State College Police with these townships). Designing any ICT-based system to be compatible with all these disparate agencies is a remarkably complex task; one that grows more even complex when the trying to incorporate federal agencies into the mix.

Even though these and other issues have resulted in ICT implementations that often fail to completely realize the promised gains, and often lead to new challenges, individuals in law enforcement continue to perceive ICTs as capable of benefiting law enforcement activities [7, 8, 17, 33]. So, law enforcement agencies and policymakers continue to look for opportunities to exploit ICTs to improve services. The impetus to develop integrated criminal justice information systems is rooted both in the failure to effectively share information within and across organizational boundaries and in the continued perception of ICTs ability to improve policing..

2.2 Integrated Criminal Justice Systems

Integrated criminal justice information systems (ICJS) encompass technological infrastructures, governance policies, and work practices and procedures intended to facilitate effective communication and sharing of information both within and across organizational and jurisdictional boundaries. Projected benefits of using ICJS include improved data quality, time and money savings, timely access to information, improved safety, greater efficiency and information sharing [12, 16]. Some have posited second-order benefits to ICJS use such as deterrence as a result of a perception that law enforcement is more knowledgeable of who commits crimes [4].

Claims that the benefits of ICJSs are now well-established [11, 12] are contested by empirical studies that make clear many of the benefits of ICJS initiatives are still unrealized [13, 15]. For example, the National Conference of State Legislatures (NCSL), a major advocate of the deployment of ICJSs, states explicitly that agencies should not expect to realize a financial savings as a result of their ICJS initiatives [24]. This echoes findings form

studies of ICJS initiatives in law enforcement that note efficiency gains are often offset by the costs of the increased resources required to support the ICT [28].

Echoing these findings the recent research provides inconclusive evidence as to the impact on efficiency from using ICJSs. As noted, efficiency in one area of law enforcement has in some cases been offset by the decrease in efficiency in others. The Charlotte-Mecklenburg Police Department (CMPD) found that as a result of the implementation of their KBCOPS system, officers were now spending on average 30 minutes to two hours keying paper-based incident reports into a web-form for submission to the system [39]. CMPD effectively transferred the burden of data entry from record-management staffers to field officers. Instead of achieving the goal of improved data collection, CMPD found that officers would skip fields that lead to increased data entry.

In their study of a prototype wireless system as part of Pennsylvania's JNET system, Sawyer et al. [33] reports similar findings. They found that connectivity drops, dual-layer authentication, and battery drain all added significantly to the cost to the officer trying to use the system. These issues were exacerbated by the limited IT support available within the agencies studied. Additionally, they report that the implementation of wireless access did not alter existing organizational structures. Officers using the technology still relied on existing communications structures (dispatch).

The issues that helped spur the movement to integrate ICTs in law enforcement agencies have also plagued the development efforts. The National Association of Chief Information Officers (NASCIO) found that aging and often incompatible infrastructures; a limited and fragmented communications spectrum; and stovepipe development practices hamper development efforts [26]. Similarly, the General Accountability Office in a report on the Department of Homeland Security's (DHS) Project SAFECOM, found the goal of enhanced interagency communication to be hampered by limited standards, lack of funding, and a lack of interagency collaboration [15]. Battles over organizational turf continue to be a major obstacle resulting in a “lack of resource pooling, lack of information sharing, poor procedure development, and a lack of adaptation [10, 16]”. Implementing the needed institutional reforms has often been relegated a low priority or resisted altogether [10].

In spite of the documented difficulties and ambiguous results, agencies continue to press on with their ICJS initiatives, and new initiatives continue to proliferate both in the United States and globally [27]. Each of these initiatives has its own design methodology, governance structures, and system components [24]. The move is to integrate disparate information systems: but, the level of integration has been in many ways limited to the scope of the project. It would be too easy, and sad, if this were allowed to negate the overall goal of nationally integrated systems.

Integrating criminal justice systems has become one of the more visible aspects of digital governance. It seems wise, if not imperative, that the “best” design, development, and deployment practices are identified early in the process so that designers, managers, and policymakers can make informed decisions regarding their initiatives. This is why we focus our research on empirical and conceptual insights on the design of ICJS.

3. ARJIS STUDY BACKGROUND

Two practical motivations motivate our research on ARJIS. One is the increased public interest in the development of systems that allow disparate law enforcement agencies to communicate and coordinate across jurisdictional and organizational lines. Individual law enforcement agencies are complex and sophisticated organizations and as a result, any effort to integrate multiple law enforcement agencies must necessarily be complex and sophisticated as well. Our research seeks to understand how this organizational complexity and sophistication impacts system development and use.

A second motivation is that system development efforts at a macro level have, to date, been largely ad hoc. Funding and direction has come both from the top in the form of grants and directives from federal agencies such as the Department of Homeland Security and the Department of Justice (National Institutes of Justice and Office of Justice Programs), as well as from the bottom through the efforts of individual officers and units. System design choices have varied across initiatives in terms of technologies, policies, and governance structures. We seek to identify the successful design choices and practices with the goal of contributing to a more “standardized” approach to ICJS development.

We are also motivated to inform theory by drawing from our results on the nature and structure of the interdependencies among technical architectures and public-sector organizational governance. Specifically, we seek to understand how institutional influences impact the development, operation, and governance of integrated criminal justice systems. By doing so we intend to engage and extend a body of knowledge that is vibrant, discordant, and often not well-developed theoretically [34].

3.1 Social Informatics

We take a Social Informatics perspective grounded in Institutional Theory to this research. Social Informatics focuses on “the design, uses, and consequences of ICTs (information and communications technologies) that takes into account their interaction with institutional and cultural contexts [19, 32].” Our view of ICTs is that they are embedded with social context, shape human social context, and are shaped by human social context. From this perspective, neither the technology nor the user is without agency. Nor is the user an abstraction that engages the ICT in a social vacuum. Rather, both the technology and the user are embedded in a highly complex and evolving context that is constituted and shaped by both.

Social Informatics research engages a range of social theories, drawing from a range of viable theoretical positions that engage human activity as bound by social norms and organizational constraints. Social informatics begins with the premise the people are social actors. That is, people's individual agency (their ability to act) is constrained by a number of social forces.

For this research we draw on Institutional Theory as our theoretical framework. Institutional Theory posits that social organizations are comprised of rules, sanctions and physical structures embedded in social and cultural contexts [5]. Social institutions constrain and define social life either through coercion (regulative), through the definition of appropriateness (normative), or by example (cognitive) [6, 18]. These social institutions can and often do transcend organizational boundaries. Examples of institutions include professional associations, corporations, political parties, bodies of professional knowledge, governments, cultural practices, and even technologies. These institutions constrain and shape the way organizations behave.

Seen as a means to explain how longstanding sets of formal, informal and overlapping social forces and organizations interact, Institutional Theory is most appropriate to the law enforcement domain. Law enforcement agencies are highly social and cultural agencies, with strong professional associations and well-defined bodies of knowledge. They are embedded in a complex system of institutions that includes government and community as well as the historical, the political, and the technological. We seek to understand to understand how these institutions drive and constrain the design, development, and use of ICJSs.

3.2 Research Approach

The practical value of studying ICJS makes these worthy objects of study. Conceptually, these systems are an ideal example to study inter-organizational systems. Such systems are complex due to the relationships among technical elements (their computing architecture) and the institutional structures in and across which they exist and influence (the broad range of agencies, levels of government and stakeholders in and out of the public sector) [27]. The nature and effects of the relations among technical architectures and institutional structures links various social science and computing research areas, and the theorizing in this area, while nascent, is quite active [20, 22, 27, 31].

Since this research is part of a larger, comparative, study a common framework is critical. The common framework we used builds on that reported on in Sawyer, et al. [33] and focuses attention to:

- Computing infrastructure elements to include nature and structure of wired and wireless connection, throughput, coverage, reliability, and costs.
- The types, uses and characteristics of the devices being used.

- The functionality, feature sets, design principles, and development efforts regarding applications and systems software. This includes attending to issues with security and authentication.

- Information sharing, uptake, distribution, and needs. This includes sources of information, cross-system and cross-boundary information sharing, and the volume, types, and uses of information

- Work activities of stakeholders from both task analysis and work structuring perspectives. This includes a range of stakeholders (such as mobile and fixed-location users, dispatch, developer, administrators, etc.) and a range of work environments.

- Governance structures and processes. This includes both operational governance (of the work being done and of the systems development efforts) and inter-organizational governance (problem-resolution, policy-setting and decision-making).

This research framework also guides the cross-time (temporal) nature of our data collection. We used five forms of data collection. Three focus on gathering primary data: interviews (face-to-face, by phone, and via email, depending on the point of the interaction), ride-alongs with – and other direct observation of – users. We also gathered secondary documents such as reports, memos and locally-relevant material (we, of course, have done and continue to do extensive web and library research to support the field work) as well as data about device uses, data transmission, and ARJIS usage via unobtrusive means (such browser logs, server logs, and telecom activity logs).

Data from the sources are transcribed into digital format or collected at source in digital format. Data from the usage logs came in digital format. This supported our analysis across different data sets and data collection approaches. To do this analysis we are using traditional qualitative/case study data analysis approaches (see [23]). In particular, we are focusing on three techniques: interim analysis of the data to guide data collection and interpretation in the future, explanatory event matrices, and content analysis of the interview/focus group transcripts and field notes.

We are currently completing the case study of ARJIS. When the study is complete we expect to have fifteen (30 hours) interviews, six officer ride-alongs, and analysis of over 650 pages of documents. At the end of this research we expect to have a comprehensive and in-depth understanding of the ARJIS system both technologically and institutionally.

3.3 Automated Regional Justice Information System (ARJIS)

The Automated Regional Justice Information System (ARJIS) of San Diego, California is one of the preeminent criminal justice information systems initiatives in the United States. Initially a mainframe records management system accessible by multiple jurisdictions in the San Diego area, ARJIS has evolved over the past 20 years both organizationally and technologically. Organizationally ARJIS has become its own

organization embedded in the county government structure. Technologically ARJIS is in the process of developing wireless communications systems, global query application, and public safety cable television channel.

Beyond its established record of success, ARJIS is an ideal system and organization to study is that is both horizontally and vertically multi-jurisdictional. ARJIS is horizontally jurisdiction-spanning because it (the organization and the system) spans numerous local jurisdictions such as the San Diego and Carlsbad Police Departments among many others. Vertical jurisdiction spanning results from ARJIS' spanning of multiple of government including the San Diego Sheriff's Office (county), the California Highway Patrol (state), and the U.S. Border Patrol (federal) [35]. Over ten law enforcement agencies with over 10,000 law enforcement officers policing a population of over 38 million citizens are participants in the ARJIS system [9].

Technologically, ARJIS is equally robust. The ARJIS system includes over 2,500 workstations and printers, and 10,000 registered users. Over 35,000 transactions accessing 2.9 million recorded incidents, 5 million digital photos, and 4.4 million map and crime statistics occur daily. With its sheer scope, ARJIS provides a unique opportunity to study an ICJS initiative in an institutionally complex environment.

The ARJIS organization is currently engaged in a variety of different development initiatives. These initiatives include both the development of hardware and software. The ARJIS system is being developed along two separate but parallel paths. One path is the development of a web-based interface to a SQL database designed to first interact with, and then later replace the legacy systems that comprise ARJIS today. This system, called Global Query, makes use of both proprietary and commercial-off-the-shelf (COTS) applications. One benefit of these applications is that it provides a vehicle for the development of other projects such as wireless PDA access to the system for the BorderSafe project.

The other development path is more ad hoc and is driven by customer (agency) demand. These initiatives consist primarily of adding functionality or access to the existing ARJIS infrastructure. For example, the addition of "Wants and Warrants" information to the system, mapping or geographical information systems applications, and access to the databases tracking pawn shop transactions. These applications are incorporated ARJIS in a manner that is consistent with their overall design approach ensuring that everything done is towards the larger goal of integration.

4. EMERGENT DESIGN APPROACH

We observe that the ARJIS approach to designing both their information system and their organization differ from other ICJS design approaches, and most of the recommendations to be found in the literature. Gil-Garcia et al. [16] identify three types of ICJS development approaches: selective, comprehensive and

incremental. A selective approach is organizational or functional-area specific, such as a computer-aided-dispatch (CAD) system. A comprehensive approach is multi-organizational, multi-level, and is completed in a short time frame such as the implementation of a juvenile records system accessible by law enforcement, the court system, and social services. This is often driven by a comprehensive master (or enterprise) plan. Like the comprehensive, the incremental approach is multi-level and multi-organizational. However functionality is added incrementally relative to a comprehensive enterprise plan.

These three approaches adhere to the literature on ICJS development that recommends two fundamental components of the ICJS process. The first is an extensive, detailed project plan / design document itemizing not only the technical specifications, but the organizational design and the information sharing procedures and policies [16]. The second views significant government support and leadership and infrastructure upgrading as critical to ICJS development efforts [24]. These two views are consistent with what Truex et al. [36] identify as assumptions and ideas that have been privileged in systems development discourse in general.

Based on our examination of ARJIS, we suggest a fourth approach that we call "organic." The organic approach appears similar to the incremental approach, but is also "bottom-up," or a combination of both bottom-up and top-down. Instead of being driven by an advocate in a position of high authority and guided by a structured and enterprise-level project plan, the "organic" approach is emergent [22]. This approach to development is closer to approaches Truex et al. [36] identify as marginalized in the literature, yet as they also point out, it is also likely to be closer to how system development actually occurs.

By emergent we mean that the organic approach reflects what Truex and colleagues argue is a more realistic (and thus viable) basis for designing, developing and delivering information systems [36, 37]. Truex et al. [37] identify principles that define both emergent systems and the organizations that develop them (see Table 1).

Table 1 Emergent Systems Principles

Emergent Systems Principles	Meaning
Always analysis	Organizations, and thus, their systems, are dynamic. This demands that analysis must continually engage these changes.
Incomplete and usefully ambiguous specifications	Specifications can never be complete, but can serve as guiding principles and to establish parameters
Continuous redevelopment	The system is ever evolving, never complete and uses are always changing.

Adaptability orientation	Developers, leaderships and users engage these changes as matters of course, not as exceptions or problems.
Back-channel communications	Interactions among key stakeholders must include many informal and formal mechanisms, and be seen as a discourse, not as a set of contracts.
Emergent IT organization	The symbioses among the system and the systems development organization suggests that both must be flexible, changing dynamically and often in relations to one-another
Proper reward systems	Developers, users and leadership must be incentivized to engage in this dynamic approach to systems.

5. CONCLUSIONS AND FUTURE DIRECTIONS

This section is incomplete and tentative, as befits the interim nature of what we are reporting. Having said this, we are beginning to see two second-order effects resulting from organic approach, as we discuss briefly. In the discussion in the previous section, we began developing the idea of organic development as a strategically sound, technologically wise, and operationally useful approach. Here we elaborate on that by invoking the architectural metaphor of the New England farmhouse.

The first second-order benefit we can highlight is that the bottom-up and top-down means of organic development are mutually reinforcing. For example, we find that when ARJIS' staff engages the individually requested components in the system, it provides them the opportunity to bring those systems into adherence with ARJIS data standards. ARJIS accomplishes this by requiring agencies to comply with ARJIS standards in order to have their data incorporated into the larger system.

We also find that this attention to engaging stakeholders at the work level creates a virtuous feedback loop to the leaders involved in governance. They are more willing to engage in long term thinking and supporting larger-scale activities because they see these as helping achieve both strategic and tactical objectives.

The architectural metaphor represented by organic development is visible in the way New England's farmhouses have evolved. The strategic goals (house the family, access to livestock, to protection from the elements) guide the design (top-down). The evolving needs of the family (as children grew and married, as parents aged, and as the number and nature of the

family changed) led to additions, expansions and new functionality. Likewise the changing nature of the farm (more livestock, more tools) and protection from elements (connecting house to barn) led to an evolving structure. The structure is architecturally distinct and functionally sound. Each farmhouse is different, but collectively they are identified because they emerged following the same top down, while responding to similar bottom-up pressures – that vary due to specific local forces.

6. ACKNOWLEDGMENTS

This work was supported in part by National Science Foundation grant IIS-051238.

Our appreciation to the officers and staff of ARJIS whose assisted us in studying and understanding their organization.

7. REFERENCES

- [1] 9/11 Emergency Communications. Centre County Government Office of Communications ed., 2005.
- [2] Fall Enrollment Summary. The Pennsylvania State University ed. University Fact Book, 2005.
- [3] State College (borough), Pennsylvania. U.S. Census Bureau ed. State & County Quick Facts, 2003.
- [4] Agrawal, M., Rao, H.R. and Sanders, G.L. Impact of Mobile Computing Terminals in Police Work Journal of Organizational Computing & Electronic Commerce, Lawrence Erlbaum Associates, 2003, 73.
- [5] Avgerou, C. Information systems and global diversity. Oxford University Press, Oxford ; New York, 2002.
- [6] Avgerou, C., Siemer, J. and Bjorn-Andersen, N. The academic field of information systems in Europe. European Journal of Information Systems, 8 (2). 136.
- [7] Brown, M.M. The benefits and costs of information technology innovations: An empirical assessment of a local government agency. Public Performance & Management Review, 24 (4). 351.
- [8] Brown, M.M. and Brudney, J.L. Learning organizations in the public sector? A study of police agencies employing information and technology to advance knowledge. Public Administration Review, 63 (1). 30.
- [9] Bureau of Justice Statistics. Local Police Departments 2000, 2000.
- [10] Clayton, R. and Haverty, D.M. Modernizing Homeland Defense and Security. Journal of Homeland Security and Emergency Management, 2 (1). Article 7.
- [11] Dunworth, T. Criminal Justice and the IT Revolution. Policies, Processes, and Decisions of the Criminal Justice System, 3. 371-426.
- [12] Dunworth, T. Information Technology and the Criminal Justice System: A Historical Review. in Pattavina, A. ed. Information Technology and the Criminal Justice System, Sage Publications, Inc., Thousand Oaks, CA, 2005, 1-28.
- [13] General Accountability Office. Homeland Security: Agency Plans, Implementation, and Challenges Regarding the National Strategy for Homeland Security. Report to the

- Chairman, Subcommittee on National Security, Emerging Threats, and International Relations, Committee on Government Reform, House of Representatives, 2005, 198.
- [14] General Accountability Office. Information Technology: FBI Needs an Enterprise Architecture to Guide Its Modernization Activities, 2003, 30.
 - [15] General Accountability Office. Project SAFECOM: Key Cross-Agency Emergency Communications Efforts Require Stronger Collaboration, 2004, 27.
 - [16] Gil-Garcia, J.R., Schneider, C.A. and Pardo, T.A. Effective Strategies in Justice Information Integration: A Brief Current Practices Review Center for Technology in Government, Albany, NY, 2004.
 - [17] Hoey, A. Techno-Cops: Information Technology and Law Enforcement. *International Journal of Law and Information Technology*, 6 (1). 69-90.
 - [18] Hossam, A.-H. Institutional Theory. Appalachian State University and York Universite eds. Theories used in IS Research, 2005.
 - [19] Kling, R., Rosenbaum, H. and Sawyer, S. Understanding and Communicating Social Informatics: A Framework for Studying and Teaching the Human Contexts of Information and Communications Technologies. *Information Today*, Inc., Medford, New Jersey, 2005.
 - [20] Law, J. and Bijker, W.E. Technology, Stability, and Social Theory. in Bijker, W.E. ed. *Shaping technology/building society : studies in sociotechnical change*, MIT Press, Cambridge, Mass., 1992, 32-50.
 - [21] Manning, P.K. Policing contingencies. University of Chicago Press, Chicago, 2003.
 - [22] Markus, M.L. and Robey, D. Information Technology and Organizational Change: Conceptions of Causality in Theory and Research. *Management Science*, 34 (5). 583.
 - [23] Miles, M.B. and Huberman, A.M. Qualitative data analysis : a sourcebook of new methods. Sage Publications, Newbury Park, 1984.
 - [24] Morton, H. Integrated Criminal Justice Information Systems. National Conference of State Legislatures. ed., 2004.
 - [25] National Association of State Chief Information Officers (NASCIO). Concept for Operations for Integrated Justice Information Sharing Systems, 2003.
 - [26] National Association of State Chief Information Officers (NASCIO). Concept for Operations for Integrated Justice Information Sharing Systems, 2003.
 - [27] Northrop, A., Kraemer, K.L. and King, J.L. Police use of computers. *Journal of Criminal Justice*, 23 (3). 259.
 - [28] Nunn, S. Police information technology: Assessing the effects of computerization on urban police functions. *Public Administration Review*, 61 (2). 221.
 - [29] Ratcliffe, J.H. Crime Mapping and the Training Needs of Law Enforcement. *European Journal on Criminal Policy and Research*, 10 (1). 65.
 - [30] Rendell, E.G. and Miller, C.J.B. Pennsylvania State Police 2003 Annual Report. Pennsylvania State Police ed., 2003, 48.
 - [31] Rudman, W., Clarke, R. and Metzel, J. Emergency Responders: Drastically Underfunded, Dangerously Unprepared Report of an Independent Task Force Sponsored by the Council on Foreign Relations, 2003.
 - [32] Sawyer, S. Social Informatics: Overview, Principles and Opportunities. *Bulletin of the American Society for Information Science and Technology*, 31 (5). 9.
 - [33] Sawyer, S., Tapia, A., Pesheck, L. and Davenport, J. Mobility and the First Responder. *Communications of the ACM*, 47 (3). 62.
 - [34] Sawyer, S. and Tyworth, M., Integrated Criminal Justice Systems: Designing Effective Systems for Inter-Organizational Action. in Sixth Annual National Conference on Digital Government Research: Emerging Trends, (Atlanta, GA, 2005).
 - [35] Scanlon, P. ARJIS: Automated Regional Justice Information System, ARJIS, 2004, 30.
 - [36] Truex, D., Baskerville, R. and Travis, J. Amethodical systems development: the deferred meaning of systems development methods. *Accounting, Management and Information Technologies*, 10 (1). 53-79.
 - [37] Truex, D.P., Baskerville, R. and Klein, H. Growing systems in emergent organizations. *Communications of the ACM*, 42 (8). 117.
 - [38] Walker, L. Integrated Criminal Information System Trends in 2002: Communication, Collaboration, Cooperation. Courts, N.C.F.S. ed., 2002.
 - [39] Williams, S.R. and Aasheim, C. Information Technology in the Practice of Law Enforcement. *Journal of Cases on Information Technology*, 7 (1). 71.

8. ENDNOTES

1. We use the term ‘public safety’ to refer to ambulance and fire services; ‘criminal justice’ to refer to the courts, prison, and parole systems; and ‘law enforcement’ to refer to police agencies. First responders are found in both law enforcement and public safety organizations.
2. See <http://www.arjis.org/>
3. For a comprehensive history of the use of ICTs in law enforcement see Dunworth, T. *Information Technology and the Criminal Justice System: A Historical Review*. in Pattavina, A. ed. (2005) *Information Technology and the Criminal Justice System*, Sage Publications, Inc., Thousand Oaks, CA, 1-28.
4. See <http://www.usdoj.gov/jmd/mps/manual/crm.htm#content>
5. There are 2,565 individual municipalities in Pennsylvania. Additionally, Pennsylvania is a commonwealth with strong townships and relatively

weak county governments. These two factors may make Pennsylvania one of the more institutionally complex states in the Union.

6. As we complete the study, we anticipate directly mapping ARJIS activity to these principles (and in

doing this engage and contributes to institutional theory).

SESSION 4A

CRISIS MANAGEMENT 2

Moderator

Anthony Cresswell, University at Albany/SUNY, USA

Titles and Authors

TIME-CRITICAL INFORMATION SERVICES: Analysis and Workshop Findings on Technology, Organizational, and Policy Dimensions to Emergency Response and Related E-Governmental Services
Horan, Thomas A.; Marich, Michael; Schooley, Ben

Secure Interoperation for Effective Data Mining in Border Control and Homeland Security Applications
Adam, Nabil R.; Atluri, Vijayalakshmi; Koslowski, Rey; Grossman, Robert; Janeja, Vandana, P.; Warner, Janice

Project Highlights: Multiple Agency and Jurisdiction Organized Response (M.A.J.O.R.) Disaster Research (NSF award # 428216)
Batteau, Allen W.; Brandenburg, Dale; Brewster, Jon; Seeger, Matt; White, Suzanne

TIME-CRITICAL INFORMATION SERVICES

Analysis and Workshop Findings on Technology, Organizational, and Policy Dimensions to Emergency Response and Related E-Governmental Services

Thomas A. Horan, Ph.D.

Clairemont Graduate University
School of Info. Systems & Tech.
130 East Ninth, Claremont, CA
Tom.Horan@cgu.edu

Michael Marich

Clairemont Graduate University
School of Info. Systems & Tech.
130 East Ninth, Claremont, CA
Michael.Marich@cgu.edu

Ben Schooley

Clairemont Graduate University
School of Info. Systems & Tech.
130 East Ninth, Claremont, CA
Ben.Schooley@cgu.edu

ABSTRACT

This paper discusses a general framework for understanding and researching end-to-end performance of inter-organizational e-governmental services and reports the findings from an expert workshop held at the National Center for Digital Government. The focus of this paper is on time-critical information services (TCIS) – the medical necessity to deliver emergency services as rapidly as possible coupled with the dependence of these services upon accurate and timely information from multiple organizations. The authors outline a TCIS model and then discuss an invitational workshop that allowed for expert (academic and practitioner) input and feedback on TCIS dimensions and the best means for understanding their occurrence in on-the-ground emergency services. Workshop participants analyzed TCIS from a socio-technical perspective and provided conceptual, practitioner and methodological critiques and suggestions. Overall, participants found the concept of TCIS to be a valid model for understanding, researching, and developing e-government systems within the specific context of emergency response as well as within the broader context of time-critical services to the public. Workshop recommendations focused on the need to closely assess inter-agency and inter-organizational *information exchanges* along and between three levels: technical, organizational, and governance. The paper concludes with a discussion about future research directions based on the analytical framework and workshop findings.

Categories and Subject Descriptors

H.4 [Information Systems Applications]

Keywords

e-government, emergency medical services and response, performance evaluation, time-critical information services

1. INTRODUCTION

This paper provides an overview of time-critical information services (TCIS), with specific reference to its use in emergency response and related e-governmental services. The paper is composed of three parts. The first part provides a preliminary review of the concepts and methods relating to the TCIS concept, including recent work by the authors on the concept. The second part contains a summary of the expert workshop that was conducted to explore the concept of TCIS. The final part provides a summary and direction based on these two activities.

2. TIME-CRITICAL INFORMATION SERVICES

This section examines the concepts and methods relating to the TCIS, placing the research within the context of e-government. The interaction between technology, the people and organizations, and the policies governing organizations and their usage of information technology (IT) is then explained. Our recent work on the concept is then presented.

2.1 Taking Just-in-Time to e-Government

From an e-government perspective, contemporary Emergency Medical Services (EMS) systems such as 9-1-1 are emblematic of what could be considered a “time-critical information service”. The time-critical element refers to the medical necessity to deliver emergency services in as rapid a time period as possible, executed through a chain of dispatchers and responders. The information-critical element refers to the fact that this service has become highly dependent upon information—from the nature and location of the incident, to the medical needs of the patient that should be attended to at the awaiting hospital [1], [6], [21]. Moreover, both time and information service elements are fundamentally organizational issues: effective and timely service depends upon all participating organizations working cooperatively and utilizing information technology effectively [14]. In particular, wireless

carriers, emergency dispatch center administrators (e.g., Public Safety Answering Points), law enforcement, fire and EMS officials, and state and local political leaders need to cooperate to deliver an integrated set of 9-1-1 services [12], [13], [16].

Of course, time efficiency and effectiveness are not new concepts to business, computer, and information professionals. Concepts such as Just-in-Time (JIT) and Business Process Reengineering (BPR) have become central to private sector business operations and information technology planning. These concepts, and others like them, coincide with the notion of improving the integrated delivery of business related processes and services using information technology. Real-time, information rich, computer aided processes to reduce costs and increase efficiencies in the value chain are driven by the private sector axiom that “time is money”.

While private sector oriented information systems have focused on the critical role of information technology in achieving JIT delivery and improved value chain management, our thesis is that similar attention is needed to those public sector services that are also highly time and information dependent. EMS represents an illustrative application domain of JIT in public services, where in this case “time is lives”.

Transportation research, specifically Intelligent Transportation Systems (ITS) research, has looked at how IT can create efficiencies (i.e., time and cost savings) in transportation related public services [8]. However, to date, limited research has been done to understand how IT can enhance time-critical functions for public services (such as EMS) that are information intensive. Further, we have found a paucity of e-government research that investigates time-critical dimensions to those public services that depend on multiple cooperating organizations.

2.2 The Need for a Socio-Technical Approach

Before launching into specific dimensions of our model, it is perhaps useful to explain the socio-technical orientation of our research. Important dimensions related to performance improvement (including time related dimensions) are the overarching organizational, institutional, and technical systems that interact with each other [3]. Improving timeliness and overall performance means looking at the supporting technology and how it interacts with the people and organizations using them, as well as the policies governing organizations and their usage of IT.

In her private sector (business) research, Markus [15] explains the need for more integration between IT development and organizational change management; a need that she has termed “technochange.” Markus states, “Technochange situations call for big improvements in organizational performance. These improvements cannot happen unless tasks, jobs, and organizational processes all change along with IT (p. 7).” Similarly, but from a public sector research perspective, Fountain [3] explains the need for coordination between inter-organizational networks of people, organizations, and policies to gain a better understanding about how they interact with IT.

Fountain’s Technology Enactment Framework provides a set of guiding propositions that explain the interactions between objective information technologies, organizational forms, and institutional arrangements. These interactions influence the design, perceptions, and uses of information technologies.

A socio-technical approach must be holistic, addressing the entire system. Fountain [5] explains that for a large socio-technical system, the mere accumulation of more sophisticated technology and specialists is insufficient. In order for agencies to develop new, integrated e-government programs, they must first examine the entire existing socio-technical system to assess their organizations’ readiness to integrate digital government – from technical, managerial, and political perspectives.

Time-critical public service systems (such as EMS) are large, complex, and often dynamic, and thus need to be investigated from such a perspective. Sussman [20] and colleagues speak of “Complex, Large-Scale, Integrated, Open Systems” (CLIOS). A significant aspect of their analysis is examining the nesting of technological systems within institutional processes and linkages. To restate in the domain of e-government, a CLIOS (such as EMS) includes a technical system that is nested in a social and institutional system. The degree and nature of these linkages are complex, and in this sense complex includes dynamic, emerging, and not fully predictable elements. Moreover, their (CLIOS) approach suggests the utility of portraying a complex system as an important conceptual step toward understanding how the system operates and evolves. Included in this portrayal is the need to examine links across the various institutions, including their complex inter-organizational socio-technical permutations.

While there is a need to understand how a complex inter-organizational system operates and interacts with its various parts, there is also a need to understand the best methodological approaches for examining this phenomenon. The primary purpose of the time-critical information services symposium is to better understand the socio-technical nature of such systems as well as the best methodological approaches for examining them. The motivation for this inquiry is that a grounded conceptualization will provide a means to analyze and improve time-critical services to the public. More specifically, it is aimed to understand the impact that timely information can have on governmental processes, service performance, and on the general public welfare. The conceptual emphasis is on the interplay between information, information technologies, the organizational network charged with delivering EMS, and the policies that govern EMS agencies and service provision. It is structured as an inter-disciplinary investigation at the interface between information science and organizational science, focusing on the information and organizational processes involved in time-critical information services.

This research builds on an initial investigation of EMS conducted under sponsorship from the Minnesota Department of Transportation and extends the analysis to additional regions so as to inform national e-government, highway safety, healthcare, and emergency management policies.

2.3 Our Conceptual Model

For the last three years the authors have examined time-critical information services within the context of rural emergency response [10]. Our analysis has led us to identify several features that we think are important for understanding end-to-end performance in time-critical information services. We provide a preliminary conceptual illustration in Figure 1. Based on this preliminary analysis there are several components that would enter into a conceptual model for time-critical information services, both in regard to EMS specifically and other public services generally. These components include: 1) the time and information critical elements of the service, 2) inter-organizational linkages that include both qualitative organizational elements as well as “hard” information flow elements, 3) end-to-end elements that consider performance metrics within and across the process flow, and 4) context variation elements such as normal versus peak conditions (in terms of service demand). We summarize this preliminary conceptual model below, then describe the symposium topics, discussion, and findings, and finally discuss future research directions.

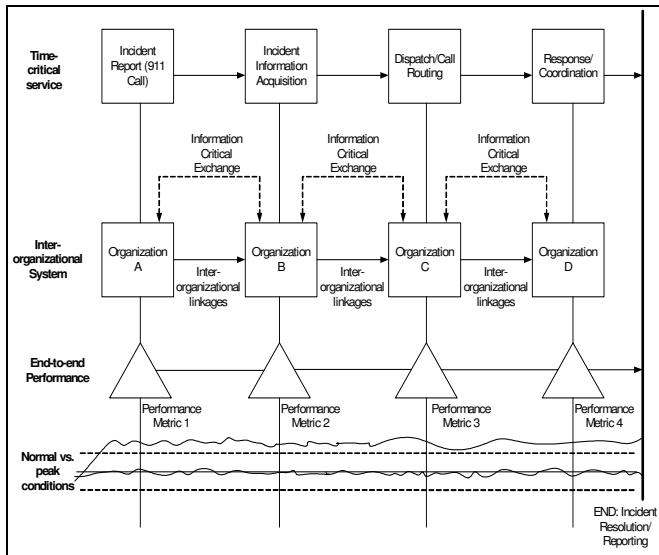


Figure 1. Illustration of TCIS Concepts

2.3.1 Time-Criticality

The first concept focuses on the time-criticality of the e-governmental services and improvements therein. In the case of EMS, time is measured in minutes and seconds and these differences can have a pronounced impact on the health condition and survivability of patients. As illustrated by the top row of Figure 1, a point of departure for examining EMS is the linear sequence of events. The “end-to-end” EMS service typically begins with a consumer action (placing the call), involves the private sector (the cellular service provider) delivering the call, the public sector (PSAP or state police) receiving and dispatching the call, the private and/or public sector (an ambulance service or fire/police) providing first response, transport and health care

services, and finally, either a public or private sector hospital delivering additional health care services. It is this end-to-end process that needs to be made efficient. The business sector has integrated end-to-end time-critical concepts such as Business Process Reengineering (BPR), Just-in-Time (JIT) information systems, and Supply Chain Management (SCM) into day-to-day operations.

This leads to the notion of how organizations create processes and cultures that recognize and embrace time-criticality and improvements therein. That is, it is plausible that organizations can to a greater or lesser extent embrace the concept of ‘time-criticality’ in their processes, procedures, and preferences.

2.3.2 Information Exchange

Even in the early days of the 9-1-1 system, information was important. Conveying voice information on location and the nature of an emergency proved valuable to response agencies to help prepare for the unique circumstances of each individual emergency incident. As this information becomes digitized, more information can be made available to responders. The second concept focuses on the dynamics surrounding the exchange and sharing of information throughout the end-to-end time-critical process. One of the key problems of governmental services is that information travels serially and sequentially, from one processing unit to the next, often with time-consuming feedback loops when incomplete or inaccurate information is detected. In the case of EMS, information mistakes can be tragic, as when an ambulance is unable to find the victim due to inaccurate location information about the scene of the accident. Further, new technologies, such as computer-aided dispatch, can establish technical inter-organizational linkages, but do not ensure a comprehensive linkage between organizations to accommodate the dynamic human response elements inherent in responding to unexpected events [22].

In our research, we found that while the service hand-offs were incrementally improving the time and information flows, what was missing was an integrated “organizational awareness” between organizations on how they were to inter-operate [9]. The software can only go so far without the “org-ware”—that is, integrating organizational policies across the service (see Figure 1, second row from the top).

2.3.3 Inter-organizational Systems (IOS)

A time-critical information service is typically inter-organizational in nature – there is a hand off from one agency to another, or if within an agency, across the functions of an agency. The design of an information system for a service delivery system is as much of an inter-organizational phenomenon as it is a technical undertaking. Businesses have extended the aforementioned BPR concept to include a network of organizations, across value chains, to improve business processes with partners, vendors, suppliers, distributors, and sales channels. The key lesson from two decades of BPR experience, however, is that BPR necessitates organizational understanding and change [7]. This private sector concept of the need for organizational change,

induced by information technology implementation, has recently been expanded into public sector e-government research [17].

This third concept focuses in on the formal and informal organizational relationships, as well as how they affect and are affected by information exchange strategies. New technologies can be developed and implemented to enhance organizational performance, but this raises the question about how well these technologies and systems affect the inter-organizational linkages and relationships.

2.3.4 End-to-End Performance

For time-critical information services, end-to-end performance is what matters to an injured citizen (see Figure 1, third row from the top). For example, it makes little difference for an operator to dispatch quickly if the ambulance takes a very long time to arrive and/or goes to the wrong location. The critical descriptor here is “end-to-end”. Measuring effectiveness across organizations (end-to-end performance) is essential to understanding how public services are delivered to the public, the level of service (timeliness, quality) with which they are delivered, and how the network can be improved to deliver better services in an information-critical and time-critical manner. The challenge is how to implement this “end-to-end” concept within and across emergency provider organizations.

While information systems are often implemented to address separate silos of a governmental process, the end-to-end nature of time-critical information services facilitates or at least allows for information systems that can report on overall system performance. The real challenge then becomes working across very different organizational cultures—for example, departments of transportation, law enforcement, and fire/ambulance response—to achieve a holistic understanding of the total service performance. The concept of “end-to-end” highlights this aspect of e-government service delivery.

2.3.5 Normal and Extreme Events

The end-to-end performance of a time-critical service is not only a function of system processes but also a function of exogenous occurrences that tend to pressure or stress service delivery. Extreme events, such as storms, natural disasters, or terrorist attacks provide obvious examples of situations that could cause a range of system failures including overload or even collapse. This raises research issues about how information systems could be used to enhance the ability of first responders (e.g., EMS) to ramp-up service capabilities rapidly and effectively despite these exogenous occurrences.

Research focused solely on a system’s normal behavior cannot fully characterize the full range of possible system dynamics. In [4] it is explained that an inter-organizational “ecology”, or dynamic inter-organizational information sharing system, must examine “punctuating events, processes of co-evolution, and other dynamics that are not often captured in research studies (p. 2782).” Thus, from a research perspective, extreme events

provide unique opportunities to examine the systems, including the functioning of cooperative organizations under “extreme accident” conditions, such as system overload or collapse. From a systems performance perspective, dealing with crises or extreme events can be a pivotal test of the overall system management capability [11].

3. WORKSHOP STRUCTURE

The TCIS Workshop was structured to provide an opportunity for expert (academic and practitioner) input and feedback on the TCIS dimensions outlined above and the best means for understanding their occurrence in on-the-ground services in the EMS area. The workshop featured experienced academics, researchers, and practitioners and was conducted in April 2005 at Harvard University in cooperation with the National Center for Digital Government. Additional co-sponsorship was provided by the State and Local Policy Program, Humphrey Institute, University of Minnesota in collaboration with the ITS Institute, University of Minnesota

The first level of discussion was at the overall conceptual level, for as [19], note: “Perhaps the best way IS [Information Systems] professionals can facilitate the validation of conceptual models is to generate high-quality conceptual models at the outset” (p.89). Moreover, attributes outlined by these authors, in particular the extent to which they are accurate and complete, in addition to high-level conceptual development with both stakeholder and empirical validation was followed in this exploratory examination and workshop.

3.1 Conceptual Overview and Discussion

Dr. Horan opened the symposium with an overview of the workshop objectives and agenda. He first described the topic, time-critical information services as an emerging concept that combines the e-governmental aspects of service delivery, but does so under time constraints. He introduced the conceptual framework that had guided his team’s research on the topic (see Figure 1). This conceptualization stressed four dimensions: 1) time-criticality; 2) information exchange; 3) inter-organizational systems performance; and 4) normal and extreme events. Dr. Horan then reviewed the substantive focus on EMS as an exemplar of time-critical information services, making note of the longer response times and higher fatality rates in rural areas. He then laid out the objectives of the symposium, in terms of reviewing, discussing, and critiquing both the concepts of time-critical information specifically in EMS as well as more broadly in e-governmental research.

Jane Fountain, Associate Professor of Public Policy, Harvard University, and Director of the National Center for Digital Government, provided a broad policy and conceptual introduction and a review. Dr. Fountain focused much of her comments on the distribution of power within and across government agencies, as well as how time-critical information services, and emergency medical services, relate to cyberinfrastructure concepts. Fountain mentioned The Atkins Report, which describes a “cyberinfrastructure” where databases and centers communicate

with one another. Sensors, tools, and information systems are used to provide performance information to an organization, inter-organizationally, and across a network. She related the cyberinfrastructure discussion to EMS, where such technologies could have a strong impact on the service. Dr. Fountain then turned her attention to the concept of governance. Though technologies could provide an important impact on a government service, she noted that EMS demonstrates a need for strong governance. This is due to the lack of shared, or network, accountability for service outcomes. Dr. Fountain suggested that a model be used to demonstrate extreme, or stress situation cases, focusing on who is in charge of the incident (and thus document that there is no end-to-end accountability). She also mentioned that it is critical to build tools for end-to-end usage rather than for one “leg” of a service.

Joe Sussman, Professor at the Massachusetts Institute of Technology, conducts research into the planning, investment analysis, operations, management, design, and maintenance of Complex, Large-Scale, Integrated, Open Systems (CLIOS). Dr. Sussman has applied CLIOS concepts to many applications, both within the United States and abroad. He has had significant involvement in the Intelligent Transportation Systems (ITS), where he has assisted in guiding the U.S. national program. In his presentation, Dr. Sussman gave two examples that he believed were appropriate for making a comparison to systems that supply emergency medical services. The first example was that of the transportation and environment system in Mexico City, Mexico. In explaining how the CLIOS process was used in this case, he stated that it exhibits features of *nested complexity* and *evaluative complexity*. The second example was that of transporting spent nuclear fuel from many nuclear power plants within the United States to a central site in Yucca Mountain, Nevada. Dr. Sussman noted that this case also demonstrated the complexities that were involved complex systems performance and assessment. And with regard to ITS, he noted that these systems are also complex and involve use of established technologies in the areas of communications, control, electronics, and computer hardware and software. A key issue that these two cases and ITS bring to EMS, according to Sussman, is the need to understand the way that organizations will interact with one another and the changes that must be made to these institutions to successfully deploy complex technologies.

3.2 Methods and Impacts

This section of the workshop turned to the state of practice with an eye to the approaches and methods for determining the performance of TCIS systems. Two complementary perspectives were offered. First, David Aylward, Director of ComCARE Alliance, provided an expert perspective from his involvement in EMS nationally. Second, Lynne Markus, Professor at Bentley College, provided a perspective informed by her long-standing research on inter-organizational systems.

The ComCARE Alliance is an organization whose mission is to improve emergency response through the innovative use of information technology. The ComCARE Alliance is a seven year old national coalition of 100 organizations that represent nurses,

doctors, emergency medical technicians, wireless companies, public safety and health officials, transportation companies, automobile companies and safety groups, and others who are working to educate the public and policymakers. These groups are ultimately responsible for the 7,000 9-1-1 systems nationwide, as well as related emergency communications systems. Mr. Aylward, Director of ComCARE, began his presentation by stating that emergency response has missed the information revolution, leaving all groups (such as emergency medical technicians, nurses, doctors, wireless companies, and public health and safety groups) within the system disconnected. Mr. Aylward believes that the use of voice alone is not sufficient for emergency response crews and a perfect device or system that can be put in place to provide all of the voice and data information needs does not yet exist. He pointed out that automobile crashes are only a small part of emergency management and would therefore be insufficient to drive changes. Rather, he argues that the underlying medical condition of the patient is paramount. Mr. Aylward believes that in order to improve EMS, several items must be examined: facilitation services research, metrics collection, and a roll-over system for 9-1-1 calls. Facilitation services, as described by Mr. Aylward, consist of security, rights management, and directory services. With regards to metrics, he concluded that knowledge needs to be measured, since FARS data is insufficient. He observes that the lack of a roll-over system for 9-1-1 calls creates a stressful situation for small 3-person PSAPs.

Lynne Markus discussed her research on inter-organizational systems and its implications for time-critical information services generally and emergency response services, specifically. She noted that while most of her research has been in the private sector, such as enterprise business systems, the lessons apply to the public sector to a large extent. She outlined three major spheres in which this research has occurred: 1) adoption of technology, 2) how innovation of technology is used, and 3) the outcomes. These three spheres capture the sequence of activities and processes that occur when new systems are introduced. As TCIS is outcome based, Dr. Markus highlighted two possible outcomes – (1) service provisions, such as quality and cost and decision making (how the knowledge will be used) and (2) information processing, which includes such things as administration costs, errors in handling information, and delays. The issue of outcomes raises several questions: What role does IT play in these outcomes? How will technology be used (rather than concerned with the technology itself)?

A key point raised by Dr. Markus is the distribution of cost and benefits across systems that are implementing systems: *Who pays for the information and how do they benefit?* Her research has found this to be a particularly important, yet vexing, issue in inter-organizational systems. If there is not a mutuality of benefits to costs, then participation can be uneven and that is quite problematic for an inter-organizational system like emergency response systems. From an academic perspective, the research issue is how to analyze the distribution of costs and benefits across participants. From a policy or practical perspective, the issue is how to accommodate a diverse arrangement that can ensure a mutuality of costs and benefits.

Finally, Dr. Markus concurred with others, who had noted the need to look at multiple levels of inter-organizational systems. It is one level (and an important level) to look at service cooperation, but to fully understand that level, one needs to also look at the overall institutional architecture.

3.3 Case Applications in the United States

This session of the symposium provided a “grounded” perspective on how time-critical information services are being realized within the context of specific EMS systems. The presenters addressed both rural and urban perspectives, including the role that technological improvements have played in enhancing performance and related organizational and policy challenges to achieving service improvements. The workshop provided an opportunity to obtain feedback about “on-the-ground” challenges to innovative technology use. Presenters during this session of the conference included Kevin McGinnis, an EMS system builder, Barbara Pletz, EMS Director of San Mateo County, and Bradley Estothen, Engineer and Project Manager for the Minnesota Department of Transportation.

Kevin McGinnis began studying EMS systems in 1974, and has been an EMS system builder ever since. His experience in EMS is extensive and includes working as a paramedic, an EMS trainer, a regional EMS coordinator, a hospital emergency department director, Maine's state EMS director for 10 years, an EMS system consultant, serving the NHTSA conducting statewide EMS evaluations, a Program Advisor for the National Association of State EMS Directors, a director of a hospital-based ambulance system, and serving as the Maine EMS trauma system development coordinator. Mr. McGinnis discussed the state of rural emergency response as virtually unchanged over the last 30 years, involving voice communications and manual (as well as some automated) data collection. He presented arguments for the need to integrate data communications into the EMS system, particularly for rural areas where emergency response times can be over an hour. For example, data communications could be used for EMS technicians to communicate with physicians in a life threatening situation when out in a rural area. Currently, data communications are not prevalent in rural EMS and there is much work still to be done. He discussed how interdisciplinary operations, or operations involving multiple agencies and organizations, cease in major events (such as 9/11). Mr. McGinnis discussed how emergency technicians do not want a lot of information pushed to them, that they have high-stress jobs, and do not want to become overwhelmed by excessive data feeds (and information overload). Technicians work in highly dynamic environments where every emergency situation differs. This environment is better suited to having access to a variety of information that they could pull when they need it. Mr. McGinnis provided a preliminary display of what he believes to be an effective user interface for EMS personnel. This hypothetical user interface would provide quick links to real-time information as needed by technicians, such as hospital emergency room wait times, number of available beds, types of physicians on duty at awaiting hospitals, availability of medications, and so forth. Such

a user interface does not yet exist, nor does the underlying information system to support it.

Barbara Pletz has been the San Mateo County EMS Program Administrator since 1988. She is a registered nurse with over 30 years of California EMS experience. She is an active participant at the state level on various EMS committees and is former president of the EMS Administrators' Association of California where she served as its legislative chair for six years. Ms. Pletz described how San Mateo County uses EMS response time data. She explained that this data is used to determine how their overall system is performing and to determine the level of service that their sub-contracted service providers are delivering. Since the county sub-contracts EMS to American Medical Response (AMR), they closely track and observe performance data to ensure that the ambulance service meets performance benchmarks set forth in their contract. Ms. Pletz provided a specific example of how the data was used. Ambulance response times were increasing in one particular region (the area is divided into 5 regions). An inquiry into why response times were increasing revealed that AMR had reduced staffing levels in that region, which had negatively impacted their ability to respond in a timely manner. According to agreements set forth in their contract, AMR was fined by San Mateo County and AMR staffing levels were subsequently increased. According to contract, AMR reports quarterly on data related to responses, transports, costs, revenue, earnings, average patient charge, and collection rates. In addition, AMR is subject to additional audits by the county. Any incident responses that do not meet the time frame benchmarks indicated in the contract must be tagged with a reason why. Reasons include: bad address, change in priority level, traffic, weather conditions, cancelled call, and others.

Currently, real-time performance information is not collected. Performance is monitored on a quarterly basis. Ms. Pletz discussed some of the lessons they have learned through implementing technology in the field to enhance EMS. She described the county's attempt to implement hand-held computers to enter and communicate emergency incident data in the field. She explained that the implementation failed. However, a significant amount of electronic data is used to augment voice. This data is entered into on-board PC's (first responders) and laptops (ambulances) and transmitted over the web via wireless transmission. A single electronic patient record is created at the time of a CAD dispatch. Some of the CAD data is included along with first responder and ambulance paramedic patient care record data and emergency department outcome data (with working diagnosis and disposition). The data is available for managers to query, slice, and create reports. Ms. Pletz explained how much of their system-wide success has not come due to specific technologies, but due to standardization. Performance tracking beyond the initial EMS response is limited. The number of patient diversion hours is tracked and shared with other hospitals and is currently the only method to monitor individual hospital performance.

Bradley Estothen is an engineer and project manager in the Minnesota Department of Transportation's Intelligent Transportation System department. In his presentation, Mr. Estothen reported on the many ways that the General Motors

OnStar system is impacting emergency services. He described OnStar as a vehicle safety and security system that effectively provides emergency assistance. Mr. Estochen stated that there are plans to standardize the OnStar system in 2006 vehicles. In the event of an automobile accident, the OnStar system shares information such as the location of the vehicle, the final resting position, the time of the accident, how the report was initiated (such as air bag deployment), and the type of vehicle. However, no personal information is shared by the OnStar system. Mr. Estochen stated that one concern that emergency response service personnel believe they face, as a result of systems such as OnStar, is the risk of homeostasis (where risk-takers feel a greater sense of security due to the monitors and may take even greater risks). In his duties as the Mayday Field Operation Test program manager, Mr. Estochen works with the United States Department of Transportation, the Mayo Clinic, OnStar, 9-1-1 service providers, and technology consultants. He stated that the Mayo Clinic has a desire to improve the outcome for patients involved in vehicle accidents. Mr. Estochen briefly explained the value of the Condition Acquisition and Reporting System (CARS), whose features allow transportation personnel to enter, display, and disseminate travel, road, weather, and traffic information. However, he noted that the CARS, while currently available in ten states, is being adopted slowly.

3.4 Research Needs and Policy Implications

This fourth session consisted of an interactive discussion among workshop participants focusing on the research and policy directions suggested by prior case and research presentations and discussions. This segment of the workshop was moderated by Lee Munnich, a Senior Fellow at the University of Minnesota. It provided the invited speakers a chance to offer additional commentary and it gave the workshop participants the opportunity to provide relevant feedback based on their expertise and professional experience. The following paragraphs provide some highlights of this insightful discussion.

Dr. Fountain made two points, one regarding the role of people and the other regarding issues of legality. First, she pointed out that in building the systems that support EMS, the people that have to work with these systems (the users) need to be considered, as well as the patients (the customers). Second, the legal obligations associated with the private systems, such as OnStar, need to be examined within the context of the legal obligations of the larger public service (EMS).

The costs associated with any type of public EMS system are of vital concern. Mr. Alyward's comments stressed that designing and building information systems that are affordable to all local EMS systems and agencies must be considered. He added that there are many examples of the exorbitant costs associated with not doing this. Mr. Alyward also added that the commercial world is building technical devices to access a distributed service based network that would accomplish these goals. There is currently a business plan for the integration of these systems into emergency services.

Dr. Sussman mentioned the importance of maintaining archival data. He felt that this type of data would be useful for strategic planning. Ms. Flaherty, a NHTSA Program Analyst, cautioned that the mission of business is different than the mission of healthcare. She suggested that those working with EMS systems should exercise care when applying business models to EMS. Adding to the discussion of legal issues, Mr. McGinnis stated that the use of information from vehicle recorders has legal implications. He also offered that time criticality must be weighed against service outcome. Dr. Markus mentioned the value of creating a web-based forum for local EMS agencies to share their information technology and business model implementation experiences with other agencies nationwide.

Sheila Madhani, a program officer for the Institute of Medicine (IOM), National Academies of Sciences, mentioned that she is currently involved in a major study with "The Future of Emergency Care in the United States Health System". She discussed that the scale of analysis is broad, with the focus of the study extending beyond pre-hospital care and into post-arrival at a hospital, pediatric emergency care, and rehabilitation. Final reports are being compiled and are due in April 2006. Madhani mentioned that pre-hospital care is only one portion of the emergency care system and that it has been valuable for the IOM to look at pre and post hospital care at multiple levels, including policy and technology, such as is the case with TCIS.

4. CONCLUSION

In sum, participants found the concept of time-critical information services to be a valid model for understanding, researching, and developing e-government systems within the specific context of emergency response as well as within the broader context of time-critical services to the public. Some participants noted that as a general conceptualization the model seemed abstract and that the case studies were crucial to providing insight and 'grounding' to the concepts. Discussions about the three levels of relationships—technical, organizational, and governance—were particularly instructive to tying operational examples to TCIS concepts. Further, the general view from participants was that the inter-organizational and "end-to-end" issues and concepts were the most practical and useful. Understanding outcomes from the 'end' is what carries significant value in terms of lives saved from more timely services, etc. Moreover, there was recognition that while both upstream and downstream activities mattered, performance in the hospital, once the accident victim arrived, was a critical aspect of the service that should be examined in the future.

Participants agreed that both quantitative and qualitative measures are valuable for evaluating TCIS dimensions. Note was made of the importance of visually-based systems, such as had been demonstrated through the Arena demonstration. Another point was made about the possibility of considering perspective-based metrics, such as from the consumer, organizational, and user perspectives. Further, there is the looming question of the distribution of costs and benefits, especially about how it related to who will be paying and who will be benefiting. There was endorsement about the need for grounded case studies to improve

understanding of these issues. Finally, the concept of evaluative complexity seemed to capture many of these issues.

Participants argued that substantial institutional fragmentation exists, making inter-organizational cooperation difficult. However, there was a strong sentiment that existing institutional fragmentation should not be treated as a given. The governance dimension precisely points at the need to consider policy level actions that could be taken at the local, regional, and national levels. One issue regarding policy is the apparent lack of competition to drive system improvements and the possible use of this, or related mechanisms, needed to inspire innovation.

Participants noted that there had been a general movement toward trying to utilize performance metrics in government, and to some extent, in e-government. In thinking about performance, participants advocated keeping the perceptions of the customer in mind. Moreover, the metrics should not just be oriented around the status quo, but to keep in mind a vision of the possible, not just existing conditions. A global view of high performance could be devised to help drive system change and improvement.

Participants described the need to identify IT solutions that could be appreciated and used at the “end user” level and within work systems. While most of the discussion had focused on the service level implications of performance metrics, participants also noted that performance metrics could be used to drive governance as well as analysis. Operational tests could also provide insight into business models for inter-organizational performance as well as technical aspects.

Returning to the simulation, the visual nature of it suggests alternative scenarios can be developed to better understand what is possible and to generate consensus around such changes. As one participant noted, “people don’t know what they don’t have”. When asked the question: *“How can TCIS advance our understanding of e-governmental services and contribute to the well being of citizens?”* participants noted the importance of having a dynamic interaction between general models of e-government, such as those suggested by TCIS, and grounded case study examples. Also, a number of social issues could be more fully examined; for example, the privacy tradeoffs (and regulations such as HIPAA) that could be involved in providing more health information electronically.

5. FUTURE DIRECTIONS

Subsequent to the research symposium, the authors have had additional discussions and interactions with the participants including responding to a Request for Information (RFI) from the National Highway Traffic Safety Administration (NHTSA) on Next Generation 9-1-1 Systems. These interactions have further confirmed the value of continuing research efforts to expand the TCIS model. As such, our future research falls into two primary areas. The first is in regards to understanding the needs of the users of TCIS systems from multiple dimensions. The second is in response to the need for a national model for building performance into EMS systems.

A key to understanding the needs of the many users of a complex inter-organizational public service is to understand the context of the existing system. This conference highlighted the need to understand the context of inter-organizational information exchange from three levels: technological, organizational, and policy. Several inter-organizational e-government researchers have also recently confirmed the value of understanding public systems from multiple perspectives [2], [18], [23]. Thus, one of the next steps for the authors is to examine information hand-offs from one emergency response organization to another across different deployments in the U.S. through a grounded case study approach. The goal would be to understand the operational system, and then to understand the context of information hand-offs from technological, organizational, and policy perspectives. Such context will provide a way to understand how performance information is acquired and shared across organizations, the barriers and synergies to sharing such information, and the requirements for overcoming barriers to include performance information tracking in an “end-to-end” system design.

Within the United States, work is currently underway within the Department of Transportation (DOT) at a national level to develop the next generation 9-1-1 (NG 9-1-1) system that integrates voice, video, and data. However, research conducted by the authors has shown that there exists a challenge to ensure that adequate information related to the end-to-end system performance is captured. Knowledge of the performance across the system will be of paramount importance not only to each of the stakeholders, but also to the organizational elements that are responsible for facilitating the service level agreements between the service providers. From this perspective, therefore, the authors plan to develop a framework for incorporating performance information into the Intelligent Transportation Systems (ITS) architecture as well as the NG 9-1-1 systems architecture and preliminary concept of operations (CONOPS) proposed by the U.S. DOT for EMS. The framework will then be applied to the ITS architecture and CONOPS document for NG 9-1-1 systems. The authors intend to provide a set of recommendations for incorporation into these systems.

6. ACKNOWLEDGEMENTS

This Workshop was sponsored by the Digital Government Program, National Science Foundation (Award no. 0508938). The workshop was conducted in cooperation with the National Center for Digital Government, Harvard University. Additional co-sponsorship was provided by the State and Local Policy Program, Humphrey Institute, University of Minnesota in collaboration with the ITS Institute, University of Minnesota. A related paper on the conceptual model presented herein is forthcoming in *Communications of the ACM*.

7. REFERENCES

- [1] Arens, Y., & Rosenbloom, P. (2002). Responding to the Unexpected. Report of the Workshop. New York, NY., March 1.
- [2] Dawes, S. and Prefontaine, L. (2003). Understanding New Models of Collaboration for Delivering

- Government Services. *Communications of the ACM*, 46(1), 40-43.
- [3] Fountain, J. (2001). *Building the Virtual State: Information Technology and Institutional Change*. Washington, D.C.: Brookings Institution Press.
- [4] Fedorowicz, J., Gogan, J., Ray, A. (2003). Interorganizational Ecology and Information Visibility. Proceedings of the Ninth Americas Conference on Information Systems, p. 2775-2783.
- [5] Fountain J. (2004, Oct.). Digital government and public health. *Preventing Chronic Disease*, 1(4). Retrieved from: http://www.cdc.gov/pcd/issues/2004/oct/04_0084.htm.
- [6] Hale, J. (1997). A Layered Communication Architecture for the Support of Crisis Response. *Journal of Management Information Systems*, 14(1), 235-255.
- [7] Hammer, M. (1997). *Beyond Reengineering: How the Process-Centered Organization is Changing Our Work and Our Lives*. New York, NY: Harper Business.
- [8] Horan, T. (2004). Information Systems to Improve Surface Transportation. In Gillen, D., and Levinson, D. eds., *Assessing the Benefits and Costs of ITS: Making the Business Case for ITS Investments*, Boston: Kluwer Academic Publishers.
- [9] Horan, T., and Schooley, B. (2005a). Time-Critical Information Services. *Communications of the ACM*. (Accepted and forthcoming).
- [10] Horan, T. and Schooley, B. (2005b). Interorganizational Emergency Medical Services: Case Study of Rural Wireless Deployment and Management. *Information Systems Frontiers*. 7(2), pp. 155-173.
- [11] Horan, T. and Sparrow, R. (2004). Managing Digital Infrastructures. In Zimmerman, R. and Horan, T. eds., *Digital Infrastructures*, London: Routledge Press.
- [12] Jackson, A. (2002). *Recommendations for ITS technology in emergency medical services* (PATH Record Number 26202): Washington DC: Medical Subcommittee of the ITS America Public Safety Advisory Group.
- [13] Lambert, T. (2000). Chapter 6: The Role of ITS in the Emergency Management and Emergency Services Community. In U. o. M. Center for Advanced Transportation Technology, ITE, USDOT, ITSA (Ed.), *Intelligent Transportation Primer*. Washington, D.C.: Institute of Transportation Engineers.
- [14] Mayer-Schonberger (2003). Emergency Communications: The Quest for Interoperability in the United States and Europe. In Howitt, A. and Pangi, R. eds., *Countering Terrorism: Dimensions of Preparedness*. Cambridge, MA: MIT Press.
- [15] Markus, M.L. (2004). Technochange Management: Using IT to Drive Organizational Change. *Journal of Information Technology*, 19(1), 4-20.
- [16] Potts, J. (2000, Summer). Wireless phone calls to 911: Steps toward a more effective system. *Currents*.
- [17] Scholl, H. (2004). Current Practices in E-Government-induced Business Process Change (BPC). Presented at the National Science Foundation Digital Government Program Annual Meeting. Retrieved from: <http://dgrc.org/dgo2004/disc/presentations/egov/scholl.pdf>
- [18] Scholl, J. (2005). [Interoperability in e-Government: More than just smart middleware](#). Paper presented at the 38th Hawaiian Conference of System Sciences (HICSS38), January 2005.
- [19] Shanks, G., Tansley, E., Weber, R. (2003). Using Ontology to Validate Conceptual Models. *Communications of the Association for Computing Machinery*, 46(10), 85-
- [20] Sussman, J. M. (2002). *Representing the transportation/environmental system in Mexico City as a CLIOS*. Paper presented at the 5th Annual US-Mexico Workshop on Air Quality, Ixtapan de la Sal, Mexico.
- [21] Turoff, M., Chumer, M., Van de Walle, B., Yao, X. (2004). The Design of a Dynamic Emergency Response Management Information System (DERMIS). *Journal of Information Technology Theory and Application*, accepted 12/25/03 and forthcoming.
- [22] Turoff, M., (2002). Past and Future Emergency Response Information Systems. *Communications of the ACM*, 45(4), pp. 29-32.
- [23] Williams, C. and Fedorowicz, J. (2005). A framework for analyzing cross-boundary e-government projects: the CapWin example. [DG.O 2005](#): 313-314.

Secure Interoperation for Effective Data Mining in Border Control and Homeland Security Applications¹

Nabil R. Adam, Vijayalakshmi Atluri, Rey Koslowski, Robert Grossman, Vandana P. Janeja, Janice Warner
{adam,atluri,vandana,jwarner}@cimic.rutgers.edu, rkoslowski@earthlink.net, grossman@uic.edu

1. PROJECT SUMMARY

Our NSF funded project aims at providing decision makers with the ability to extract and fuse information from multiple, heterogeneous sources in response to a query while operating under a decentralized security administration. Our motivation comes from US Customs, which embarked on a major modernization initiative of its Information Technology systems. Drawing in data from Customs trade systems, targeting inspectors review manifest information as well as strategic and tactical intelligence to determine “high-risk” shipments and containers. This entails a considerable level of communication and data sharing between various government agencies. Based on the idea of “Smart Borders”, the system will utilize data available from different agencies, ports and customs divisions to supplement the profiling by targeting towards anomalies, and detect various flags raised by non-conforming shipments or abnormal behavior of inbound cargos and raise a combination of alerts. The output of this project would ideally enhance the security aspect of the Automated Commercial Environment (ACE) system by incorporating the concept of semantic interoperability, anomaly detection and subsequent spatial and geographical visualization of information that can help Customs inspectors make better decisions.

2. RESEARCH ACTIVITIES

Secure Data Interoperation: We have proposed a *coalition based access control* (CBAC) model [2], extended it to *dynamic coalition-based access control* (DCBAC) model [15], which allows a user’s request to access a resource belonging to another coalition entity to be automatically translated using attributes associated with user credentials and objects. We proposed an approach to handle how local access control policies could be mapped into collaboration level policies using attributes and graphs, and examined the problem of determining appropriate attributes and their values to require from remote users in order to grant them access to a requested resource [16]. Our process transforms Role Based Access Control (RBAC) policies into attribute requirements that must be presented by external users via credentials. We extended the coalition service registry (CSR) architecture which supports DCBAC to include distributed CSR (DCSR) [14]. In a DCSR system, several service registry agents cooperate to provide controlled access to coalition resources.

Distribution of the registries results in improved availability, higher concurrency, better response times to user queries, and enhanced flexibility.

Anomaly Detection in Real-time Data Streams: We introduce an algorithm for detecting outliers on streaming data, which relies on computing a dyadic decomposition into cubes in Euclidean space [3]. If we view the dyadic decomposition as a tree with a fixed maximum size, then outliers are naturally defined by cubes containing a small number of points in the cube, or the cube itself and its neighboring cubes. The cumulative sum (CUSUM) algorithm is a standard algorithm for detecting changes in an event stream. CUSUM assumes that both normal and unusual events be modeled using standard distributions, and uses a formula based upon the log odds ratio to sound an alarm when a threshold is passed. Variations on the CUSUM also exist for cases of unknown distributions. A natural extension of CUSUM is to consider collections or ensembles of CUSUM algorithms and a rule that determines one or more members of the ensemble to apply. We consider ensembles of CUSUM algorithms defined by cells in a multidimensional data cube [4], which arise naturally when large data sets have temporal, spatial, or spatial-temporal variations. In the area of change detection in multi-modal streaming data [5], we developed a testbed containing: real time data from over 830 highway traffic sensors in the Chicago region, data about weather, and text data about events. The goal was to detect in real time interesting changes in traffic conditions. Given the size and complexity of the data, we built a separate baseline model for each hour in the day, for each day in the week, and for every 2 or 3 traffic sensors, resulting in over 42,000 separate baseline models. We also built a baseline engine to build the necessary baselines automatically. We modified an open source scoring engine to process in real time each new sensor reading, update the appropriate feature vectors, score the updated feature vectors using the baseline models, and send out real time alerts when deviations from the baselines were detected. The system and architecture we developed should apply more generally when there is a requirement to generate real time alerts from multiple streams of complex data. In [17], we introduce Tukey and Tukey scagnostics and develop graph theoretic methods for implementing their procedures on large data sets. An important advantage of this approach is that the visualization does not suffer the curse of dimensionality that effects many competing approaches. This approach appears to be a fruitful method for visually detecting anomalies in large dimensional data sets.

Anomaly Detection in Spatio-temporal Data: We have proposed a random walk based free-form spatial scan statistic approach for anomalous window detection [6]. A spatial scan statistic considers a scan window, and identifies anomalous windows by moving the scan window in the region. Earlier proposals suffer from two limitations: (i) They restrict the scan window to be of a regular shape (e.g., circle, rectangle, cylinder),

whereas the region of anomaly, in general, is not necessarily of a regular shape. (ii) They take into account autocorrelation among spatial data, but not spatial heterogeneity. As a result, they often result in inaccurate anomalous windows. To address these limitations, we proposed a random walk based Free-Form Spatial Scan Statistic (FS³). Application of FS³ on real datasets has shown that it can identify more refined anomalous windows with better likelihood ratio of it being an anomaly, than those identified by earlier spatial scan statistic approaches.

Semantic graph (SG) based knowledge discovery: We use SGs from a set of disparate sources and related ontologies [1]. We took a two-step approach: First, we created a refined enhanced graph by combining multiple relevant SGs and combining relevant knowledge from ontologies. This involved identifying relevant ontologies, reconciling different terminology, inferring new facts, and checking consistency of information in the SGs gathered from different sources. Second, having the enhanced and refined SG, we employed a semantics driven approach to detect patterns.

Diplomacy and politics: We examined how Canada has been using new information technologies to screen terrorists while enabling legitimate travel and trade. We interviewed Canadian Border Services Agency (CBSA) and Foreign Affairs officials in Ottawa and visited border crossings and ports from the Atlantic to the Pacific coasts, met with local CBSA managers and saw the technology at work. The report [7] was presented in Washington at a meeting attended by several Department of Homeland Security officials. Comments were given by Richard Stana, Director, Homeland Security and Justice Issues, Government Accountability Office. It was subsequently discussed in the New York Times, the Congressional Quarterly, The Atlanta Constitution, The Arizona Republic and several other newspapers. Several presentations have been made [8-13].

3. SUCCESS AND IMPACT

As a result of this project, several other collaborations have been generated between Rutgers and SAP, including the RFID, data interoperability and privacy project. This project has resulted in two on-going Ph.D. dissertations. The publications generated during the first two years of funding of this project are available: <http://cimic.rutgers.edu/~vandana/BorderControlPublications.htm>. N. Adam gave a talk at SAP Research, Karlsruhe, Germany, in Nov. 2004. During 2006, Dr. Rey Koslowski plans to extend his research to the EU, Australia and New Zealand. His focus is in technologies used at ports of entry to screen passengers and cargo, biometrics-based registered traveler programs, advanced passenger information systems, RFID-enabled biometric visas and ICAO-compliant “e-passports.” He will work on EU implementation of EURODAC, the asylum seeker fingerprint database, deployment of the Schengen Information System (SIS) and SISII (information system for sharing data among Schengen Convention signatory states for checking of border crossers), the European Visa Identification System (VIS). In addition to meeting with policymakers and border control officials (interior ministry, customs, immigration, foreign ministry) of EU member states, Australia and New Zealand, he plans to meet with officials at the European Commission and World Customs Organization in Brussels, the Schengen Information System secretariat in Strasbourg, the International Organization for Migration in Geneva, the Budapest Process Secretariat and Austrian Interior ministry in Vienna, Interpol in London, Europol in the Hague and the EU Agency for the Management of Operational Cooperation

at the External Borders of the Member States (Frontex) in Warsaw.

REFERENCES

- [1] N. Adam et al. "Semantic Graph based Knowledge Discovery from Heterogeneous Information Sources", Working Together: Conference on Public/Private R&D Partnerships in Homeland Security, 2005.
- [2] V. Atluri, J. Warner: Automatic Enforcement of Access Control Policies Among Dynamic Coalitions, ICDCIT 2004.
- [3] C. Gupta and R. L. Grossman, Outlier Detection in Streams With Dyadic Cubes, will be submitted to KDD 2006.
- [4] R. L. Grossman and H. V. Poor, Baselines and Change Detection Using Ensembles of CUSUM Algorithms, submitted to IEEE TKDE.
- [5] R. L. Grossman et al. Change Detection and Alerts from Highway Traffic Data, ACM/IEEE SC 2005 Conference.
- [6] V.P. Janeja and V. Atluri, "FS3 : A Random Walk based Free-Form Spatial Scan Statistic for Anomalous Window Detection". ICDM 2005, Houston,Texas, USA
- [7] R. Koslowski, Real Challenges for Virtual Borders: The Implementation of US-VISIT, Migration Policy Institute Report, June 2005.
- [8] R. Koslowski, “Virtual Borders and Homeland Security,” Beyond Terror: A New Security Agenda, Watson Institute, Brown University, June, 2005.
- [9] R. Koslowski, “Real Challenges for Virtual Borders: The Implementation of US-VISIT,” presentation and report release, Migration Policy Institute, Washington, DC June, 9, 2005
- [10] R. Koslowski, “Real Challenges for Virtual Borders,” Centre on Migration Policy and Society, Oxford University, Mar. 16, 2005.
- [11] R. Koslowski, “Toward Virtual Borders: Expanding European Border Control Policy Initiatives and Technology Implementations” An Immigration Policy of Europe? New York University and the European Union Institute, Florence, Italy, March 13-15, 2005.
- [12] R. Koslowski, “Possible Steps Towards an International Regime for Mobility and Security”, UN, Mar, 2005.
- [13] R. Koslowski, "European and Transatlantic Cooperation on Migration, Mobility and Security" Beyond the U.S. War on Terrorism: Comparing Domestic Legal Remedies to an International Dilemma, The J.B. Moore Society Spring 2005 Symposium, University of Virginia Law School, Feb, 2005.
- [14] R. Mukkamala, V. Atluri and J. Warner, A distributed Service Registry for Resource Sharing Among Ad-hoc Dynamic Coalitions, IFIP TC-11 WG 11.1 and WG 11.5 Joint Working Conference, 2005
- [15] J. Warner, V. Atluri and R. Mukkamala, A Credential-Based Approach for Facilitating Automatic Resource Sharing Among Ad-Hoc Dynamic Coalitions. *DBSec 2005*: 252-266
- [16] J. Warner, V. Atluri and R. Mukkamala, An Attribute Graph Based Approach to Map Local Access Control Policies to Credential Based Access Control Policies, ICISS 2005.
- [17] L. Wilkison, A. Anand, R. Grossman, Graph-Theoretic Seagnostics, IEEE Symposium on Information Visualization 2005.

¹Supported in part by the National Science Foundation under grant IIS-0306838.

Project Highlights: Multiple Agency and Jurisdiction Organized Response (M.A.J.O.R.) Disaster Research (NSF award # 428216)

Allen W. Batteau
Director, Inst. for Info.
Technology and Culture
Wayne State University
Detroit, Michigan 48202
313-874-7010
a.batteau@wayne.edu

Dale Brandenburg
Director, Inst. for Learning &
Performance Improvement
Wayne State University
Detroit, Michigan 48202
313-577-6674
ab3447@wayne.edu

Matt Seeger
Chair, Department of
Communication
Wayne State University
Detroit, Michigan 48202
313-577-2959
aa4331@wayne.edu

1. ABSTRACT

This paper gives recent accomplishments of the MAJOR (Multiple Agency and Jurisdiction Organized Response) Disaster Management Project (NSF #428216), including publications, model development, and field research findings.

2. PRINCIPAL INVESTIGATORS

Allen W. Batteau, PhD, Dale Brandenburg, PhD, Jon Brewster, PhD, Matt Seeger, PhD, Suzanne White, MD.

3. PROJECT OBJECTIVE

To develop an agent-based simulation of disaster response that will model both coordination and the breakdown of coordination among jurisdictions and agencies.

4. CURRENT ACTIVITIES

We have developed our basic models (the functional model of disaster response, an evaluation model of training simulators, and an agent-based model of incident response) and are now assembling case data to populate the models. We have briefed emergency services managers in Southeast Michigan with these results, and have identified specific additional areas to focus research efforts.

5. RESEARCH CONTRIBUTIONS

Although the textbook conceptualization of public administration is based on a Weberian model of bureaucracy, in federal systems (such as the US and Canada) where jurisdictions and functional

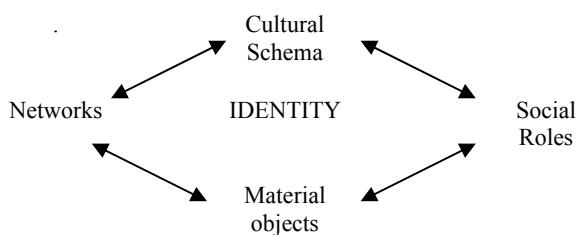
responsibilities are divided, effective response to all but the simplest, small-scale incidents requires a coordinated effort among multiple jurisdictions and agencies. We have postulated that an agent-based model provides a more realistic portrayal of this decentralized decisionmaking than does the more conventional command-and-control model. (Although the National Incident Management System, NIMS, stipulates centralized control, Hurricane Katrina clearly demonstrated the vulnerabilities of centralized schemes.)

Thus far we have had four publications and presentations resulting from this research by Klein, Brandenburg, Atas, and Maher [1], Klein, Herzog, Smolinske, and White [2], Seeger [3], and Toelken, Seeger, and Batteau [4].

6. SUCCESSES

Perhaps the greatest success of the MAJOR project to date has been the development of a dynamic model of identity that can account for coordination and its breakdown. Identity is seen as composed of four elements (cultural schema, artifacts, roles, and networks of relationships), *all of which dynamically reinforce the others*. When any of these relationships are disrupted, identity and sensemaking can collapse, reverting to a more primitive state.

Figure 1. Model of Identity



In this model the conceptual schema includes the categories of identity supplied by regional, national, or professional cultures; networks are relationships that reinforce identity; material objects are tangible symbols of identity, often with both utilitarian and symbolic value; and social roles are behaviors appropriate to identity.

The information flows between these different elements of identity take the form both of overt messages (such as would be transmitted through communication devices) and also subtle cues (a glance to a co-worker for approval). *The importance of this model is that any alteration or disruption of the information flows, such as would result from a new communication network, has the potential for challenging sensemaking and identity.*

The other significant success has been the development of methods and procedures for observing disaster response, both real and simulated. To date we have observed 11 disaster exercises, and are preparing case studies of several more. In these we have observed some of the key elements that contribute to learning (or the lack thereof) from exceptional events. These observations and case studies will be used to run our models of organizational learning and response.

An additional success has been our *Crossings* workshop, held on March 14-15, 2005, in which we explored some of the border security issues with Canadian colleagues. Although this activity was sponsored by the Canadian Embassy and a separate NSF grant (0519440), it enabled us to explore an area of cross-jurisdictional cooperation that is uniquely important to Detroit: the presence of an international border with major commercial importance.

7. CHALLENGES

Perhaps the greatest challenge of this research is identifying data resources of sufficient granularity to permit modeling of the different information flows that make up identity in real time. This is hindered by the fact that, as recently evidenced in the aftermath of Katrina, such information has the potential to embarrass public officials, particularly if there is a notable breakdown in coordination. Finding proxies for these data will remain a significant challenge.

8. PLANS

Our plans for the forthcoming year are (a) to populate our models with case data, either from first-hand observation or from after-action reports on disasters and exercises; and (b) to develop theoretical and case study articles on disaster coordination. Additionally, we are collecting data on some of the disaster coordination systems and software, especially information technology tools such as E-Team, that are currently being adopted by many emergency services agencies. These systems, which can be used both to manage disaster response and to train first responders, will alter many standard practices in emergency response; whether they will add greater resilience or brittleness to response patterns is yet to be determined.

9. REFERENCES

- [1] Klein, K. R., Brandenburg, D. C., Atas, J.G. and Maher, A. The Use of Trained Observers as an Evaluation Tool for a Multi-Hospital Bioterrorism Exercise. *Prehospital and Disaster Medicine*, 20,3 (2005), 159-163.
- [2] Klein, K.P., Herzog, Smolinske, S. C., and Suzanne White, S. Demand for Poison Control Center Services "Surged" During the 2003 Blackout. Presented at the North American Congress of Clinical Toxicology, Seattle, Washington. Also to be published in the *Journal of Clinical Toxicology*.
- [3] Seeger, Matthew W. Best Practices in Crisis communication: An Expert Panel Process. Centers for Disease Control and Prevention. 2005.
- [4] Toelken, K., Seeger, M.W., and Batteau, A. Learning and Renewal Following Threat and Crisis: The Experience of a Computer Services Firm in Response to Y2K and 9/11. *International Society for Crisis and Risk Management*. April, 2005.

SESSION 4C

E-CITIES

Moderator

Genevieve Giuliano, University of Southern California, USA

Titles and Authors

The Fully Mobile City Government Project (MCity)
Scholl, Hans J; Fidel, Raya; Mai, Jens-Erik

*UrbanSim: Interaction and Participation in Integrated Urban Land Use, Transportation,
and Environmental Modeling*
Borning, Alan; Waddell, Paul

*Simulating Impact of Light Rail on Urban Growth in Phoenix: An Application of Urbanism
Modeling Environment*
Guhathakurta, Subhrajit; Joshi, Himanshu; Konjevod, Goran; Crittenden, John; Li, Ke

Intelligent Cities
Curwell, Steve

The Fully Mobile City Government Project (MCity)

Hans (Jochen) Scholl (PI)
University of Washington
The Information School
Seattle, WA 98195-2840
+1-206-616-2543

jscholl@u.washington.edu

Raya Fidel (Co-PI)
University of Washington
The Information School
Seattle, WA 98195-2840
+1-206-543-1888

fidel@u.washington.edu

Jens-Erik Mai (Co-PI)
University of Washington
The Information School
Seattle, WA 98195-2840
+1-206-616-2541

jemai@u.washington.edu

ABSTRACT

The Fully Mobile City Government Project, also known as MCity, is an interdisciplinary research project on the premises, requirements, and effects of fully mobile, wirelessly connected applications (FMWC). The project will develop an analytical framework for interpreting the interaction and interdependence between main organizational variables and work context.

Categories and Subject Descriptors

H.1.2 [Information Systems]: Models and Principles – user/machine systems.

General Terms

Management, Performance, Design, Economics, Human Factors.

Keywords

Digital Government, mobile Government, FMWC, ubiquitous computing, pervasive computing, integration, interoperation, interfacing, e-Government, socio technical challenges

1. INTRODUCTION

Digital government holds the promise of agile, lean, accountable, and citizen-centric government operations, which are responsive, fast, effective, efficient, and sufficiently integrated [1-6]. Within this context of digital government, *mobile government*, as some refer to the use of mobile devices and applications over wireless networks for integrated voice/data communications and transactions, opens new dimensions to and avenues towards that vision. In this vein, fully mobile wirelessly connected (FMWC) applications are being examined and tested by governments for their potential in giving government field operations an unprecedented quality of immediacy in accessing information needed for critical ad-hoc decision making. Many FMWC applications are also sensitive to the ambience and to the needs of a specific individual worker. The potential utility and efficacy of these applications might significantly help advance the Digital Government (DG) agenda. Since major variables of the organizational and work context are immediately impacted, the introduction of new base technologies is highly risky. By developing an analytical and formative model, this research intends to contribute to the understanding of critical interdependencies and interactions between important variables.

2. THE MCITY PROJECT

2.1 Background

This project is being conducted with the support of the City of Seattle Public Utilities (SPU). The City of Seattle was one of the earliest adopters of DG concepts and has developed into a nationally recognized role model for innovative DG practices [7-9]. With the advent of sufficiently robust mobile Information and Communication Technology (ICT), the City began deploying and using mobile applications in its field operations. SPU was among the early adopters of first generation (1G) FMWC applications in its Water Operations group, where fieldworker crews were equipped with FMWC-enabled ruggedized laptops, cellular data phones, and Personal Digital Assistants (PDAs). Since the introduction of 1G-FMWC applications to the SPU Water Operations field crews, in 2002, numerous measured and intangible benefits have been accounted for – such as cost reductions, productivity increases, work process streamlining, increase of data integrity and quality, increased customer satisfaction, and reduced number of task delays among others [10]. These benefits are encouraging but SPU as well as City of Seattle ICT management, have become keenly aware of the multiple serious challenges when moving from a single-unit, small scale, and incremental pilot towards a multiple-agency, City-wide, ambience-specific deployment of FMWC applications with streamlined and enhanced backend interoperability as well as redesigned work processes in field operations. This study will address these challenges.

2.2 Proposed Research

Through field studies, observation and interviews at the SPU we will iteratively develop a formative model representing the constraints, interactions, and interdependencies of the studied fieldwork domains and their contexts, in which FMWC applications are either already in use or are under consideration for use. Based on this work- and task-anchored model, we will develop the requirements and characteristics of FMWC devices and applications. Finally, we will analyze and specify the organizational, social, individual, and technological impacts of FMWC applications for various fieldwork types.

3. PEOJECT OUTLINE

This project is divided into three primary task areas; Task I: preparing the research team and development of the initial model; Task II: expansion and refinement of the initial model; and Task III: completion of the model and evaluation. A timeline of three years has been scheduled to conduct the project.

3.1 Task I: Prepare Research Staff and Develop Initial Model

We are currently in the beginning stages of the research. Over the course of the coming year we will be reviewing the literature pertaining to subject areas relevant to the project. This will include FMWC applications and uses as well as literature relating to the information behavior of fieldworkers.

The SPU fieldworker population and different fieldwork types will be analyzed to determine potential participants based along previously determined selection criteria. A survey will be conducted and participants for the first round of the study will be chosen based on information obtained.

Five to seven initial cases will be observed and analyzed. Participants will undergo an entry interview. Data then collected through observation and analysis will be used to develop the initial model for the study. This stage will end with exit interviews of the participants.

3.2 Task II: Expand and Refine Models

Data will be collected from 25 additional fieldwork cases. This data will be analyzed and the findings integrated into the model. At this point we will have more than 30 cases in hand and expect general patterns of SPU fieldwork and specific FMWC requirements to emerge from the data.

3.3 Task III: Finalize and Evaluate Model

We will collect data on an additional 18 fieldwork cases. Again, we will analyze them and integrate the data into the model. The model will then be based on 50 fieldwork cases and we expect to be able to verify/falsify the patterns observed during earlier tasks. In addition, we will develop an evaluation framework for assessing the findings of the final model and expose the model to peer and practitioner scrutiny.

4. INTELLECTUAL MERIT

This project has the capacity to break new ground in the following areas: (1) With the fieldwork domain-centered approach, we go beyond specifying the factors influencing the organizational outcomes of FMWC application uses as most research has done so far, and determine the effects of their interactions and interdependencies within a formative framework. (2) With the formative framework we increase the explanatory power of the structuration perspective on ICT as introduced by Orlikowski & Robey [11]. (3) We further define the requirements for FMWC applications and workflows as well as the policy choices available to decision-makers in DG. In this regard, the proposed research contributes significantly to the understanding of the elements and processes leading to the organizational success of FMWC applications in DG, and provides a mechanism for the evaluation of such technology.

5. BROADER IMPACTS

This research will help define the strategic choices and avoid costly failures, when adopting FMWC applications in DG. It will inform both academics and practitioners about the organizational prerequisites, consequences, and the specific requirements regarding FMWC technology. The potential utility of the expected findings, however, are not limited to government. They may be highly informative to other environments.

6. ACKNOWLEDGMENTS

This project is supported through grant # IIS-0535088 under NSF program 05-551 (Information and Intelligent Systems: Advancing Collaborative and Intelligent Systems and their Societal Implications).

7. REFERENCES

- [1] Aldrich, D., Bertot, J.C. and McClure, C.R. E-Government: Initiatives, developments, and issues. *Government Information Quarterly*, 19 4 (2002), 349-355.
- [2] Bush, G.W. e-Gov: The official web site of the President's e-Government initiatives, The White House, 2002.
- [3] Osborne, D. and Gaebler T. *Reinventing government: how the entrepreneurial spirit is transforming the public sector*. Addison-Wesley Pub. Co., Reading, Mass., 1992.
- [4] Relyea, H.C. E-gov: Introduction and overview. *Government Information Quarterly*, 19 1 (2002), 9-35.
- [5] Savas, E.S. *Privatizing the public sector: how to shrink government*. Chatham House Publishers, Chatham, N.J., 1982.
- [6] Taylor, R.S. *Value-added processes in information systems*. Ablex Pub. Corp., Norwood, N.J., 1986.
- [7] Ho, A.T.-k, Reinventing local governments and the e-government initiative. *Public Administration Review* 62 (2002), 434-444.
- [8] Kaylor, C., Deshazo, R., and Van Eck, D. Gauging e-government: A report on implementing services among American cities. *Journal of Global Information Management*, 18 (2002), 293-307.
- [9] Kaylor, C. The next wave of e-Government: The challenges of data architecture. *Bulletin of the American Society for Information Science and Technology*, 31, (2005), 18-22.
- [10] Bleiler, R. SPU Technology Project Post-Implementation Review. *Public Utilities*, Seattle, 2003.
- [11] Orlikowski, W. J., & Robey, D. Information technology and structuring of organizations. *Information Systems Research*, 2, 2 (1991), 143-169.

UrbanSim: Interaction and Participation in Integrated Urban Land Use, Transportation, and Environmental Modeling

Alan Borning

Dept. of Computer Science & Engineering
University of Washington
Box 352350
Seattle, Washington 98195

borning@cs.washington.edu

Paul Waddell

Daniel J. Evans School of Public Affairs
University of Washington
Box 353055
Seattle, Washington 98195

pwaddell@u.washington.edu

1. PROJECT OVERVIEW AND IMPACTS

The process of planning and constructing a new light rail system or freeway, setting an urban growth boundary, changing tax policy, or modifying zoning and land use plans is often politically charged. Our goal in the UrbanSim project is to provide tools for stakeholders to be able to consider different scenarios, and then to evaluate these scenarios by modeling the resulting patterns of urban growth and redevelopment, of transportation usage, and of environmental impacts, over periods of 20–30 years. UrbanSim, combined with transportation models and macroeconomic inputs, performs simulations of the interactions among urban development, transportation, land use, and environmental impacts. It consists of a set of interacting component models that simulate different actors or processes within the urban environment.

2. RECENT RESEARCH ACTIVITIES

2.1 Opus and UrbanSim 4

One project this past year has been collaboratively developing a new software architecture and framework — Opus, the Open Platform for Urban Simulation — and rewriting UrbanSim in that framework. There were several factors that led us to take this step: a growing consensus among researchers in the urban modeling community that a common, open-source platform would greatly facilitate sharing systems, the desire to make the system code more accessible to domain experts, and some intractable problems with some of our previous component models (which were hard to solve due to the inaccessibility of the source code to domain experts, making rapid experimentation and testing hard).

After preliminary testing and design work that began in January 2005, we began implementing Opus and UrbanSim 4 (the latest version of the system) in March, and now have

a working version of both [4]. The system is written in Python, and makes heavy use of efficient matrix and array manipulation libraries (principally numarray). The implementation of Opus and UrbanSim 4 contains far less code than the previous implementation, yet implements a much more modular and user-extensible system, and runs faster. It also incorporates key functional extensions such as integrated model estimation and visualization.

Opus has been designed in collaboration with groups at the University of Toronto, Technical University of Berlin, and ETH, the Swiss Federal Institute of Technology in Zurich. The Toronto group has also been active in implementing a new open-source travel model in Opus; we plan to use that in our own work, both directly and to do baseline comparisons with an experimental activity-based travel model.

2.2 Statistical Analysis of Uncertainty

Predicting the future is a risky business. There are numerous, complex, and interacting sources of uncertainty in urban simulations of the sort we are developing, including measurement errors, uncertainty regarding exogenous data and other input parameters, and uncertainty arising from the model structure and from the stochastic nature of the simulation. Nevertheless, citizens and governments do have to make decisions, using the best available information. At the same time, we should represent the uncertainty in our conclusions as well as possible, both for truthfulness and as important data to assist in selecting among alternatives.

We are starting a new project to provide a principled statistical analysis of uncertainty in UrbanSim, and to portray these results in a clear and useful way to the users of the system. We are leveraging in this work a promising technique, Bayesian melding, which combines evidence and uncertainty about the inputs and outputs of a computer model to yield distributions of quantities of policy interest. From this can be derived both best estimates and statements of uncertainty about these quantities. This past year we have had some initial success in employing this technique, applying it to calibrate the model using various sources of uncertainty with an application in Eugene-Springfield, Oregon. These results are reported in a journal article recently submitted to *Transportation Research B: Methodology* [3].

2.3 Indicators and Stakeholder Interaction

Another set of activities concerns presenting the results of simulations to different stakeholders, including elected officials, members of neighborhood, business, and advocacy groups, in ways that are clear and that speak to the issues of concern for those stakeholders. Our work in this area is guided by the Value Sensitive Design methodology, an approach to the design of information systems that seeks to account for human values in a principled and comprehensive way throughout the design process.

One project involved carefully documenting and presenting the indicators that portray key results from the simulations. Our design addresses a number of challenges, including responding to the values and interests of diverse stakeholders, and balancing the value of fairness with presenting a diverse set of advocacy positions. We published the results of this work, including empirical evaluations, in the European Computer Supported Cooperative Work conference [1]. Another project has been the development of “Personal Indicators,” which distill the simulation results down in ways that speak to concerns of individual citizens. A preliminary description of this has been accepted for the ACM Computer Human Interaction Conference [2], and will form a section of Janet Davis’s forthcoming Ph.D. dissertation.

2.4 Testing Stochastic Systems

Agile software development methodologies and extensive testing have been a hallmark of our software engineering practices on UrbanSim for some years. However, we have had consistent problems adequately testing stochastic algorithms, which may give different results each time they are run. (And many of the key UrbanSim algorithms are stochastic.) We recently made major progress in this area, developing a set of design patterns for tests of stochastic systems that include distributional tests on the results of running the test repeatedly. This is supported by a sound statistical analysis of how to interpret the results from such tests, and a unit test framework that implements it. These results are currently being written up for publication.

3. COLLABORATIONS

One set of collaborations is with government planning agencies that want to apply UrbanSim to their regions. Our primary effort at present is with Puget Sound Regional Council, the metropolitan planning organization for our own region. We have also collaborated actively with MPOs in Salt Lake City, Eugene, Honolulu, Houston, and Detroit. There have also been research and pilot applications in Amsterdam, Paris, Phoenix, Tel Aviv, and Zurich. The first UrbanSim Users Group meeting in San Antonio, Texas, in January 2005, attracted some 30 participants from MPOs around the country, a number of academic researchers, and one participant from the Netherlands.

Another set of collaborations concerns the development of Opus, the Open Platform for Urban Simulation described in Section 2.1, with an emerging consortium of research teams from Canada, France, Germany, Japan, Switzerland, and the United States. We are also working with researchers at the University of Massachusetts in Amherst on the “UrbanSim Commons,” a web portal to facilitate exchange and collaboration among UrbanSim users.

4. PLANS AND CHALLENGES FOR THE COMING YEAR

We plan to release Opus and UrbanSim 4 early this year. A challenge has been balancing this constant software evolution, driven by the research agenda and problems that we encounter, with the needs of our government partners, who, after all, want a stable, working system that they can use as an ongoing part of their operational procedures. We hope that Opus will provide a workable platform for them, and are putting a great deal of our effort towards that end.

A more risky area of research will our emerging work on statistical analysis and representation of uncertainty using Bayesian melding, which is supported by a new Digital Government grant. As discussed above, our preliminary results are promising — but there are significant challenges and risks, including being able to adequately uncover the uncertainties in the input data and models, and being able to achieve satisfactory performance.

In the area of stakeholder presentation and interaction, we plan to complete the implementation and deployment of a web-based Indicator Browser, which will let interested stakeholders browse through simulation results. The choice and description of indicators can be value-laden and politically sensitive. In response to this, we have been developing Indicator Perspectives, partnering with different groups and agencies to put forth a variety of perspectives on what is most important in the results from UrbanSim, and how it should be interpreted. Our initial partners in this are Northwest Environment Watch, the King County Benchmarks Program, and the Washington Association of Realtors.

Acknowledgments

We would like to thank all of the UrbanSim research team members and collaborators. This research has been funded in part by grants from the National Science Foundation (EIA-0121326 and IIS-0534094), in part by a partnership with the Puget Sound Regional Council, and in part by gifts from IBM and Google.

5. REFERENCES

- [1] A. Borning, B. Friedman, J. Davis, and P. Lin. Informing public deliberation: Value sensitive design of indicators for a large-scale urban simulation. In *Proc. 9th European Conference on Computer-Supported Cooperative Work*, Paris, Sept. 2005.
- [2] J. Davis. Household indicators: Design to inform and engage citizens. In *CHI 2006 Work-in-Progress Papers*. ACM Press, Apr. 2006.
- [3] H. Ševčíková, A. Raftery, and P. Waddell. Assessing uncertainty in urban simulations using Bayesian melding. Submitted for publication - draft available from www.urbansim.org/papers/BMinUrbansim.pdf, 2006.
- [4] P. Waddell, H. Ševčíková, D. Socha, E. Miller, and K. Nagel. Opus: An open platform for urban simulation. Presented at the Computers in Urban Planning and Urban Management Conference, London, June 2005. Available from www.urbansim.org/papers.

Simulating Impact of Light Rail on Urban Growth in Phoenix: An application of UrbanSim Modeling Environment

Himanshu Joshi

School of Architecture
Arizona State University
Tempe, AZ, 85287
(1)-713-499-6659

Himanshu.Joshi@h-gac.com

John Crittenden

Dept. of Civil & Environmental.
Engineering,
Arizona State University
Tempe, AZ, 85287, (1)-480-965-
3420,
jcritt@asu.edu

Subhrajit Guhathakurta*;

Planning / Global Institute of
Sustainability
Arizona State University
Tempe, AZ, 85287
(1)-480-965-6343

Subhro@asu.edu

Goran Konjevod

Department of Computer
Science,
Arizona State University
Tempe, AZ, 85287
(1)-480-965-2783

konjevod@math.cmu.edu

Ke Li

Dept. of Civil & Environmental
Engineering
Arizona State University
Tempe, AZ, 85287
(1)-480-965-3171

keli@asu.edu

ABSTRACT

This paper analyzes the impact of the proposed Phoenix Light Rail transit system on future land use and household characteristics adjacent to station areas using a software-based simulation model called UrbanSim. UrbanSim is a microsimulation modeling environment, developed at the University of Washington, Seattle, which enables detailed analysis of the effect of different policy scenarios of future patterns of urban growth. Although existing literature suggests that projects such as the Phoenix Light Rail often increases the value of properties adjacent to the station areas, few studies have analyzed the nature and type of households that are drawn to these areas. The result of this analysis for Phoenix shows differential impacts of light rail on various neighborhoods with different existing land use and household characteristics. Particularly, the study draws attention to the potential gentrification of an area currently dominated by students of Arizona State University and suggests that preemptive strategies should be put in place to accommodate student housing.

Categories and Subject Descriptors

J.4 [Social and Behavior Sciences] Sociology, economics.

General Terms

Economics, Experimentation, Verification.

Keywords:

Light Rail, transit, urban modeling, ridership, simulation

1. INTRODUCTION

The task of planning urban land use and transportation systems has become more challenging in the face of an increasing set of dynamic influences on the behaviour of households and businesses. The integration of the global economy, leading to both a quantitative and qualitative change in flows of information, capital, people, and goods, has, in part, led to a realignment of land uses. While some traditional industries have become de-centered within the urban context, other forms of economic activities are leading to re-centering around new nodes. On one hand, increasing affluence among many households have increased demand for land for housing and a dispersion of households farther away from the center. On the other, households that are being marginalized by globalizing processes are finding more segregated quarters and limited ability to access urban amenities. In Phoenix, such processes are evident in the continued development of large, master-planned communities at the periphery together with the movement of jobs to new employment clusters along major transportation corridors. At the same time, downtown Phoenix is also receiving significant attention as the location for a new campus for Arizona State University and for T-Gen, the biotechnology research cluster. The adoption of a Light Rail transit system within this context has raised several concerns about its potential viability and suitability in a rapidly changing urban region.

The concern about the Phoenix Light Rail transit system cannot be addressed without knowing the future in advance. informative, each region is unique in many ways. Hence, the experiences of other regions cannot provide definite answers to questions about the suitability of the Phoenix Light Rail system within the specificity of the Phoenix metropolitan region. In this study, we adopt a more robust form of futures analysis that takes advantage of a comprehensive urban simulation model called UrbanSim (Waddell 2002; Waddell and Borning 2004; Waddell and Gudmunder 2004). Micro-simulation models such as UrbanSim allow planners to examine simulated futures based on knowledge about the behavior of various urban actors and their interactive relationships as they play out in changing the urban fabric. For

example, UrbanSim explicitly models the behavior of households, employees, and developers, as they choose their activity location based on the collective choices made by other households, employees, and developers within a given the policy environment. Each individual household and employee is a decision-making agent and is modeled separately within the simulated environment. Since UrbanSim generates overarching patterns of urban growth based on aggregation of decisions of individual “agents”, it belongs in the category of models known as “agent-based” models.

In this study, we examine the impact of the introduction of the light rail transit system in Phoenix by observing the changes in number and type of households adjacent to the proposed transit line in 2015. Given that the light rail system in Phoenix is expected to start operation in 2008, the simulated future in 2015 provides adequate time for observing long-term changes in households and land use patterns around station areas. The impact of light rail in Phoenix metropolitan region is estimated by comparing a scenario that includes light rail with the null scenario (in which light rail transit is not introduced). The results suggest that most areas adjacent to light rail stations increase in household density as we would expect based on the literature. But there are some surprising declines in household density in other areas, especially in the high density corridor next to Arizona State University. We show that such unexpected results are consistent with urban economic theory as well.

2. WHAT WE KNOW ABOUT LIGHT RAIL TRANSIT IN THE U.S.

In the later half of twentieth century, many cities in North America adopted light rail as a convenient transit system. Today more than 50 cities in the United States provide rail transit as a means of regional public transportation. There are two types of light-rail systems. The first system involves light cars, sometimes called trolleys, trams or streetcars, which run along the street and share space with motor vehicles. Such systems exist in San Diego (in part), New Orleans and Charlotte, N.C. The second light-rail system consists of multicar trains that operate along their own right of way and are separated from roadways. St. Louis; Portland, Pittsburgh; San Jose, and Buffalo all have this second type of light-rail system (Garrett, 2004). The Phoenix Light Rail system will be of this later type.

2.1 Property Values Adjacent to Transit

Economic theory suggests that accessibility afforded by public transit can add to the amenities associated with adjacent activities. For example, residents who use the transit system may enjoy reduced travel time while businesses near a transit station can face lower costs and agglomeration benefits. Thus, traditional location theory would predict that the cost benefits resulting from proximity to transit will be capitalized in the values associated with residential and business land uses (Alonso 1964; Muth 1969; Mills 1972). Some property owners may suffer a penalty from the nuisance effects of a rail system, which imparts some ambiguity to the net effect of transit proximity. However, empirical examination has shown that in a majority of the cases, residential, office, retail and industrial properties close to rail transit enjoy significant positive premiums (FTA 2000).

Studies in Boston, Philadelphia, Portland, San Francisco, Arlington/ Washington D.C., Atlanta and San Diego found that residential properties with close proximity to rail stations had

higher property values than those farther away (Landis et al. 1994, 1995; Cervero and Duncan 2002; Al-Mosaiad, et.al. 1993; Armstrong 1994; Weinstein and Clower, 1999). But higher residential property values are not apparent in San Jose, Sacramento, and Miami. These rail systems probably were not as high quality as the others or they enjoyed very low ridership. Higher system ridership tends to increase positive property premiums throughout the transit area. Light-rail transit has enhanced residential property values some 2- 18% in Portland, Sacramento, San Diego, and Santa Clara, with larger changes in cities with commuter rail systems. But not all residences benefit equally. Properties located too near a station may suffer nuisance effects, and it appears that in California the largest premiums accrue to owners of multifamily residential properties. Of the few commercial property markets studied, it appears that there are premiums of 4-30% for office, retail and industrial buildings located near rail transit in Santa Clara, Dallas, Washington, DC, Atlanta and San Francisco (Nelson 1992; Cockerill and Stanley 2002, Landis et al. 1995).

2.2 Light Rail and Urban Form

The characteristics of a region in terms of the relative patterns and layout of employment centers with respect to residential areas have been a significant factor determining transit usage (TCRP 1996). Compact cities with a dominant center leads to higher transit usage than more dispersed urban areas with multiple employment centers. All rail systems in the U.S. are focused on the CBD area and are designed to bring employees to downtown jobs (TCRP 1996, p. 5). In the case of the San Francisco Bay Area, Cervero and Landis (1992) found that although the average commuting distances did not change much when firms relocated to the suburbs, work trips by transit plummeted by a factor of almost 20. This switch from transit to auto for suburban jobs is also corroborated by studies in England (Daniels 1972) and Houston (Rice Center 1987). However, other researchers have also shown that, despite loss of CBD jobs, a significant proportion of new and relocating job centers have sought out rail transit corridors such as those along transit systems in Chicago and Washington DC (Cervero 1995, JHK and Associates 1987). Hence, concentrating both jobs and housing on rail transit corridors can substantially increase transit ridership.

2.3 The Effect of Density and Land Use Mix

There is clear evidence that increasing density and mix of land uses around transit stations lead to higher transit ridership. Several studies in the early 1990s have shown that jobs and housing tend to co-locate in order to improve employment accessibility, which, in turn, reduces congestion and improves transit ridership (Giuliano 1991; Wachs et al. 1993; Levinson and Kumar 1994; Cervero 1996; Frank and Pivo 1994). In a seminal paper examining transit demand in Portland, Oregon, authors Nelson and Nygaard note that “of 40 land use and demographic variables studied, the most significant for determining transit demand are the overall housing density per acre and the overall employment density per acre. These two variables alone predict 93 percent of the variance in transit demand among different parts of the region” (1995, p. 3-1). A research conducted for the Transit Cooperative Research Program of the Federal Transit Administration in 1996 examined data on 19 light rail transit systems and 47 commuter rail systems and concluded that station boardings (transit usage)

was positively correlated with both station area residential density and CBD employment density (TCRP 1996, p. 12).

Mixed land uses also encourage transit usage but its effect is noted as being less significant than density (TCRP 1996). Several studies on suburban activity centres have shown that a balance of employment and housing in the area causes higher transit usage with corresponding reductions in auto trips (Cervero 1989, 1996; Hooper et al., 1989; Nowlan and Stewart 1991; Ewing 1995). Other researchers have also found similar positive impacts on transit ridership when number of retail jobs in zone is considered (1000 Friends of Oregon 1993, 1994, 1995). Empirical estimates of land use mix and transit ridership connection from Seattle, Washington (Frank 1994) and San Francisco bay Area (TCRP 1996) also come to similar conclusions. Therefore, planning for complementary and mixed land uses around station areas have become accepted practice in all metropolitan areas with transit systems.

3. THE URBANSIM IMPLEMENTATION FOR MARICOPA COUNTY

Several state-of-art modeling approaches have been incorporated into the design of the UrbanSim system. Urbansim simulates the behaviour of individual agents like households, businesses, developers, and governments (as policy inputs) and their interactions in the real estate market. By focusing on the principal agents in urban markets and the choices they make about location and development, the model deals directly with behavior that planners, policy makers, and the public can easily understand and analyze. The structure allows users to incorporate policies explicitly and to evaluate their effects.

UrbanSim is not a single model. It is an urban simulation system, which consists of a family of models interacting with each other, not directly, but through a common database. There are seven different models within Urbansim (economic transition model, demographic transition model, employment and household mobility models, employment and household location choice models, household mobility model and the real estate development model). A more detailed description of each of these models and their underlying theories are available at <http://www.urbansim.org/index.shtml>. Here we focus on the specific implementation of UrbanSim for Maricopa County.

3.1 The Database

The input data included in the data store consist of parcel file

information from county assessor's office, employment data from Maricopa Association of Governments (MAG), census data, detailed land use and land valuation data, boundary layers showing environmental, political and planning boundaries also from MAG, and a chronological list of development events. A set of software tools such as ArcGIS and MySQL was used to extract the data from input files, calculate values and construct the model database in the specified format.

The data store contains all households in Maricopa County starting with the base year 1990. Each household is a separate entry in the households table with associated characteristics such as: household income, size, age of head of household, presence and number of children, number of workers, and the number of cars. In addition, the data store contains every job present in the Maricopa County by location (i.e. grid id), job sector and whether the job is home based or not. Altogether UrbanSim requires about 60 data tables which are used in the complete database. Each table has a well defined structure. The model components include a script to check consistency of tables, which when run provides warnings and error messages if any table is incorrectly formatted for UrbanSim.

3.2 Creating the Database for Running UrbanSim

The process of creating the Maricopa County database for UrbanSim required the following steps:

1. Define project boundary (Maricopa County)
2. Define the base year for data (we used 1990)
3. Generate grid (9511 grid cells, 1mile by 1mile each were generated for Maricopa County)
4. Assign unique IDs to grid
5. GIS Overlays: Parcels on grid; transportation analysis zones (TAZ) on grid
6. Allocate parcel characteristics to grid
7. Assign employment to grid
8. Reconcile non-residential space and jobs
9. Synthesize households and locate them by Grid ID
10. Generate diagnostics and resolve inconsistencies
11. Assign development types
12. Convert environmental features to grid
13. Convert planning boundaries to grid
14. Load database into MySQL
15. Run consistency checker

In this paper, we do not attempt to describe each of these steps in

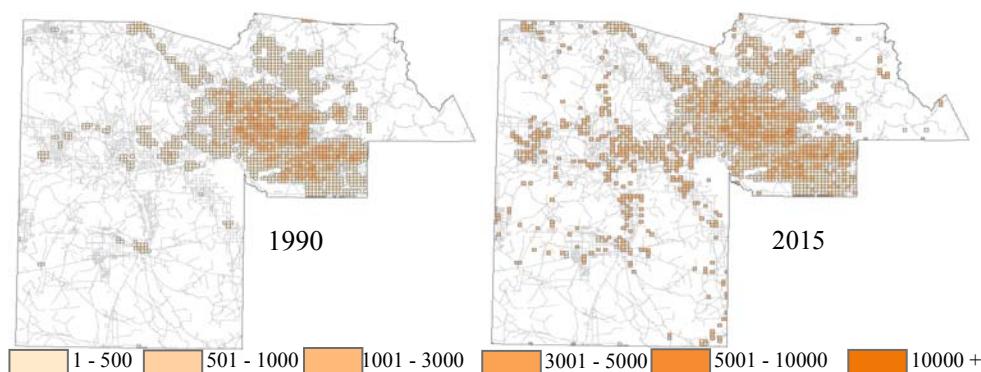


Figure 1 Household simulation using “business as usual” scenario.

detail but the steps are well documented in the UrbanSim manual available at www.urbansim.org. There is however one step, step 9 above, that requires special attention given that it is an extraordinary process when compared to most land use projection models. As mentioned earlier, Urbansim database requires information for every household in Maricopa County by their special attributes. There is no single source from which, this entire data can be obtained. For this reason, Urbansim provides a utility called as Household Synthesis Utility. As its name suggests this utility synthesizes the household data with the help of an *iterative proportional fitting algorithm*. The utility synthesizes households separately by family type for each Public Use Micro Area (PUMA) at the level of the block groups. Data sources required for the household synthesis utility are: 1) Sample of households by age of head, income, race, workers, number of children, and number of cars for families as well as non-families at the level of PUMAs from 5% Public-Use Micro-data; and 2) Block group level data for the marginal distribution tables from US Census Summary Tape file STF-3A. The algorithm iteratively matches the marginal totals at the level of block groups to varying sets of households represented in the 5% PUMS sample. When a selected set of households match closely the aggregate block group statistics, that household set is assumed to belong in that block group. Subsequently, households in the block groups are associated with the grids-IDs. In this manner the households table is populated and the synthetic process of household allocation closely approximates actual household locations.

Another important aspect of this modeling approach is the use of accessibilities as a critical driver of jobs and household locations. The information about trips between transport analysis zones (TAZs) in Maricopa County at various points in time was obtained from Maricopa Association of Governments (MAG). This data allowed us to calculate logsums by travel mode from which accessibilities were derived for incorporation in to UrbanSim data store. UrbanSim is usually run in tandem with an external travel model, so that the accessibilities can be updated at regular intervals. For our purposes we used three sets of accessibilities (1993, 1998, and 2008) based on the output from travel model used by MAG.

3.3 Model estimation

Given that UrbanSim is actually a group of models that communicate with each other through a data store, the estimation process involves separate calibration of each individual model. Most of the models are of a “discrete choice” nature and are estimated through nested logit regressions (e.g., household location choice model, developer choice model, and employment location choice model). The land price model is different from the previous set since it is the only model that is estimated through a linear regression procedure. The model parameters are derived with the help of external packages such as Limdep and SPSS.

Using the estimated parameters, two model configuration tables were generated for each model -- the model specification table, and the model coefficient table. These tables are usually the last tables to be generated. Once these tables are populated with appropriate parameters, UrbanSim model runs can be accomplished.

Figure 1 provides the 1990 and 2015 household location results for one UrbanSim run using the “business as usual” scenario.

3.4 Model Validation

Model validation is a crucial process for building confidence in the modeling results. For this paper, UrbanSim model is run from 1990 through 2000 and the simulated results are compared to the observed data to check the validity of the model. Practical constraints on creation of historical data for use in validation often preclude the feasibility of historical validation of this sort, but this remains one of the most informative ways to assess the model before putting it into operational use (Waddell and Ulfarsson, 2004). The simulation results are compared to observed data at two units of geography. As seen in Table 1, the correlation between the simulated and observed is close to 80% at the level of the grid cell. However, this correlation is lower when a larger unit of geography such as the transportation analysis zone is considered.

Table 1 Correlation of Simulated to Observed 2000 Values

	Cell	TAZ
Employment	0.8	0.71
Households	0.76	0.66
Housing Units	0.79	0.64

4. SCENARIO ANALYSIS: IMPACT OF LIGHT RAIL ON THE PHOENIX METROPOLITAN AREA

The Central Phoenix / East Valley Light Rail Transit Project, which is now under construction, will provide a convenient and comfortable transportation between Phoenix's central business district, the Sky Harbor International Airport, Arizona State University, several community college campuses, and event venues that currently draw about 12 million people each year from the region. The first phase of the project will include a 20.3-mile line that connects significant destinations in three cities – Phoenix,



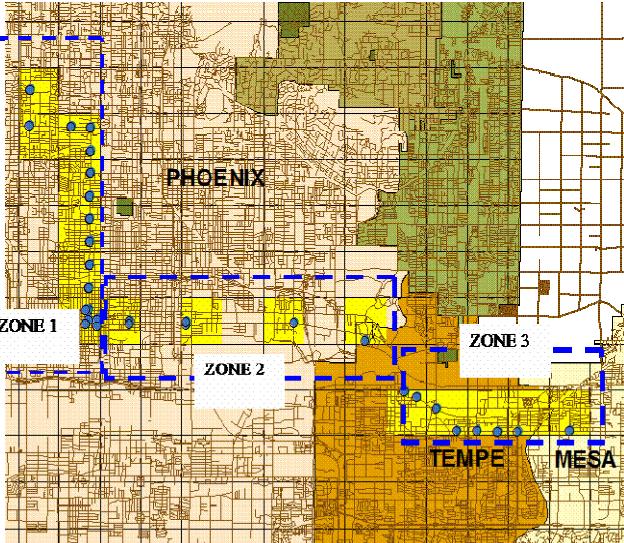


Figure 3 Delineation of three zones for Phoenix light rail study.

Tempe, and Mesa. It is expected that this phase of the project will be completed by 2008. In light of the new transportation option that will become a reality in less than three years, planners in the three cities are actively engaged in planning and redesigning the areas around the transit stops. The scenarios tested in this study takes into account many of the planned interventions around the transit stations mostly in terms of introducing mixed-use and higher density developments. Figure 2 shows a map of the overall planned system in relation to the various cities in the Phoenix metropolitan area.

The first phase of the Phoenix Light Rail project will include 32 transit stations within the cities of Phoenix, Tempe, and Mesa. These station areas are shown in Figure 3. Given that transit stops are designed to be closer together than the 1-mile grid used in the UrbanSim model, we have allocated three analysis zones each having distinct characteristics. Zone 1 radiates north from Phoenix downtown and includes most of the Phoenix downtown business district and the uptown arts district. This region includes some of the oldest neighborhoods in this metropolitan region and a fairly large downtown core. Zone 2 is currently a low density corridor that is adjacent to the commercial airport and includes many industries that have located to take advantage of proximity to the airport. This corridor also includes several low-income neighborhoods and areas with high concentration of minorities. Zone 3 is dominated by Arizona State University and activities supporting the university clientele. The concentration of student housing is high in this area. This zone also includes several ethnic retail establishments catering to a large international student community attending Arizona State University. The following analysis compares the transition of households in the three delineated zones based on scenarios with and without light rail transit for year 2015. As mentioned earlier, the scenario for different levels of transit usage was generated by changing modal split for all the TAZs that include the 32 stations mentioned earlier were changed. Accessibilities were recalculated such that for the 5% scenario, 5% of the total number of trips was added to transit and subtracted from auto. Similar procedure of increasing transit ridership was adapted for 15% and 25% scenarios. These scenarios were tested against ‘no build’, where light rail is not built and the existing mode split continues into the future. Also,

cities will be rezoning the station areas for high density, mixed-use developments. To account for this land use change, development types of the gridcells falling under stations have been changed to high density and mixed use development type. The particular light rail scenario discussed below assumes the mid-range of the three scenarios tested, that is, 15 percent of trips to and from the areas adjacent to light rail will be on the proposed Phoenix Light Rail system.

5. ANALYSIS OF URBANSIM SCENARIOS WITH AND WITHOUT LIGHT RAIL

The introduction of light rail in Phoenix metropolitan area seems to increase the number of households in zones 1 and 2 when compared to a future without light rail. Between 2008 and 2015 the number of households in zones 1 and 2 increased 19 percent and 15 percent respectively without light rail. Zone 3 also registers increase in the number of households in this scenario by 6 percent. In contrast, the scenario with light rail assigns very slight changes to household numbers in zone 1, but significant increases in zone 2. The number of zone 2 households increases by 12 percentage points during that same period when compared to no light rail scenario. Figures 4 shows the change in households by year for the two scenarios described above.

A surprising result is noticed for zone 3, which includes large concentrations of high density student housing. The scenario with light rail seems to decrease the number of households in the seven years after the commencement of light rail in Phoenix metropolitan area. Compared to the “no build” scenario, the introduction of light rail results in a decline of households by over 50 percent in Zone 3. Although the reduction in household density seems surprising, the model is behaving as expected given that capitalization of the amenity provided by light rail transit in home values may perhaps lead to new up-market developments that pushes out the lower-income student population and makes room for higher income families who prefer slightly larger quarters. This projected household transition becomes even more apparent when we examine the type of households who would prefer living adjacent to transit stops as predicted by UrbanSim.

5.1 Household transition due to light rail

Characteristics of households in the three zones show different trends based on scenarios with and without the introduction of light rail transit in 2008. In this paper we report on two of the important characteristics of projected future households adjacent to light rail station areas – income and race.

The three zones delineated for the study include, on average, low-to moderate-income households and the “no build” scenario does not change that overall character. Under the “no build” scenario, zone 1 registers the highest income of the three zones during the period of projection. Zone 2 remains the lowest in terms of average household incomes of the three zones. Both zones 1 and 2 show slight declines in real average incomes over the 7-year period of projection. Zone 3, however, registers significant decline in real average household income of about 8 percent during this period. This result changes dramatically in the scenario with light rail, especially for zone 3.

The scenario with light rail has significant yet differential impacts on zone 1 and zone 3. Households in zone 2, in contrast, are less likely to be of a different income group with or without light rail.

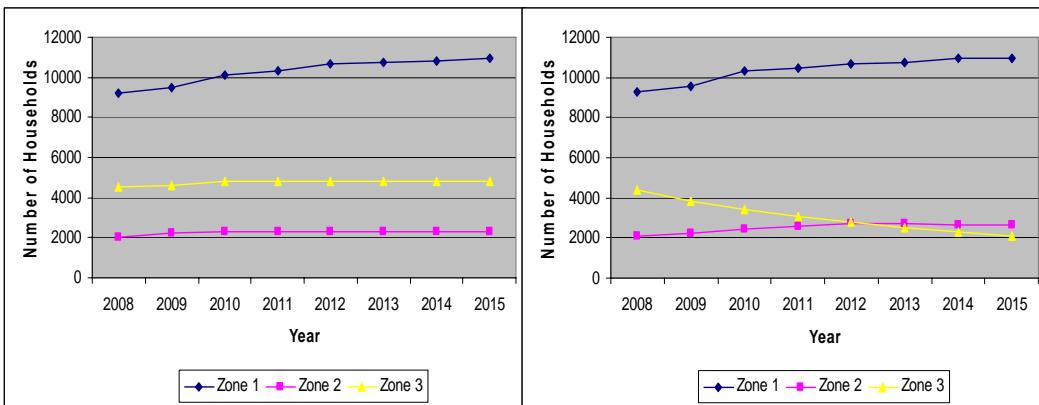


Figure 4 Change in number of households by zone without (left) and with (right) light rail.

Average household income in zone 1 is projected to decline significantly in the seven years after the initiation of light rail transit. In contrast, zone 3, which includes the student community around Arizona State University, is projected to scale up in average income levels during the same period. While households in zone 1 remain the highest in average incomes of the three zones without light rail, they give up that top position to households in zone 3 when light rail is introduced. Zone 2 households remain at the bottom in average income in either scenario.

With changes in household incomes, the racial ethnic composition of the households in the three zones also changes depending upon the introduction of light rail. In all but one scenario, the percentage White households (as determined by the racial attribute of the head of household) decline from 2008 to 2015. White households comprise about 72 percent of all households in Zones 1 and 2, and 64 percent in Zone 3 in 2008. In the scenario without light rail, the highest decline in the percentage of White households is in Zone 2 (6 percentage points) followed by Zone 1 (3 percentage points) and Zone 3 (1 percentage point). This decline is almost entirely at the expense of percentage growth of households in the “other” racial category. The “other” category is a residual category used in US Census for those individuals who do not choose among the dominant racial categories for various reasons including unwillingness to disclose or being of mixed races.

The racial make-up of the three zones seems to be very different in the scenario with light rail than the previous scenario. The decline of White households in Zone 1 is now more pronounced (10 percentage points). However in Zone 2, which had the largest decline of White households in the previous scenario, the percentage of White households now decline by only 4 percentage points. More surprisingly, percentage of White households in Zone 3 actually trend up in the scenario with light rail by a significant 6 percentage points. In essence, Zone 3 will be the most impacted area with the introduction of light rail partly due to gentrification.

5.2 The final analysis

The scenarios evaluated to test the impact of light rail on adjacent neighbourhoods in Phoenix metropolitan area show different impacts in different zones. The findings are mostly in line with the literature on transit and land use connections but also add some surprising caveats to this literature. While, as expected, Zones 1 and 2 register slight increases in residential density over 7 years since introducing light rail, household densities in Zone 3

actually decline under this scenario. This result can be explained in light of current characteristics of Zone 3 and its unique location. The household density in Zone 3 is already among the highest in the state and includes a high percentage of student households. Given the income profile of this young student population, the housing available is mostly rental, aimed at low- to mid- market clients. In addition, this area is among the most

“jobs rich” areas in the state being close to the fourth largest university in the US and to downtown Tempe. Therefore, the perceived accessibility of this area is already high and the introduction of light rail transit provides the additional amenity that would make it more desirable to up-market clients.

The projected gentrification of Zone 3 is especially unwelcome for the student population who would be gradually pushed out to areas farther from the university. Given this possible scenario, both the city of Tempe and Arizona State University will have to plan ahead for more affordable student housing in the future. The university has already embarked on an extended plan to increase on-campus student housing. The city also needs to closely monitor land use changes and real estate values in Zone 3 and look for innovative approaches for developing as well as keeping affordable housing. Regardless, this area seems to be ripe for redevelopment and the introduction of light rail will perhaps jump start the process.

An important caveat to keep in mind is that simulation models are useful tools for understanding the interaction of contextual elements and decision-agents but they are limited in their capacity to anticipate processes that have no antecedents. This limitation is more pronounced in very long-range projections. The simulation results reported in this paper is well within the period in which projections can be justifiably made, given well verified models. However, the results should be treated as informational and not definitive since human social behaviour changes through time due to adaptation and learning. Regardless, planning for future requires us to anticipate it and careful use of simulation and / or modeling tools are indispensable for this endeavour.

6. ACKNOWLEDGEMENTS

This study would not have been possible without the assistance of Maricopa Association of Governments (MAG) and we are grateful for their support with data and critique. The authors of this study would also like to thank Professors Charles Redman and Jonathan Fink at Arizona State University, for providing critical logistical and financial support for this project. Ray Quay and T. Velu also provided invaluable assistance in advising and in setting up the modeling environment. Additional funding was provided by a grant from National Science Foundation (#SBE/SES-0438447).

7. REFERENCES

- [1] 1000 Friends of Oregon. (1993). *LUTRAQ: The Pedestrian Environment Model Modifications*. Vol. 4A.

- [2] 1000 Friends of Oregon. (1994). *LUTRAQ: Model Modifications*, Volume 4. Portland, OR.
- [3] 1000 Friends of Oregon. (1995). LUTRAQ: Vol. 6, Portland OR. 112 (1986) pp. 77–87.
- [4] Al-Mosaiad, Musaad A., and Kenneth J. Dueker, and James G. Strathman. (1993). Light-Rail Transit Stations and Property Values: A Hedonic Price Approach. *Transportation Research Record* 1400: 90 – 94.
- [5] Alonso, W. (1964). Location and Land Use. Cambridge, MA: Harvard University Press.
- [6] Armstrong, Robert J. (1994). Impacts of Commuter Rail Service as Relected in Single-Family Residential Property Values. Preprint, Transportation Research Board, 73rd Annual Meeting.
- [7] Cervero, R. (1989). *America's Suburban Activity Centers: The Land Use-Transportation Link*. Boston, Unwin-Hyman
- [8] Cervero, R. (1995) *BART@20: Land Use and Development Impacts*. Institute of Urban and Regional Development, University of California at Berkeley.
- [9] Cervero, R. (1996). Jobs-housing balance revisited: trends and impacts in the San Francisco Bay Area, *Journal of the American Planning Association*, 62, pp. 492-511
- [10] Cervero, R. and J. Landis. (1992). "Suburbanization of Jobs and the Journey To Work: A Submarket Analysis of Commuting in the San Francisco Bay Area." *Journal of Advanced Transportation*, 26, 3 275–297.
- [11] Cervero, R. and M. Duncan (2002) Benefits of Proximity to Rail on Housing Markets, *Journal of Public Transportation*, Vol. 5, No. 1:1-18.
- [12] Cockerill, L. and D. Stanley. (2002). How Will the Centerline Affect Property Values in Orange County? Fullerton: California State University at Fullerton Institute of Economic and Environmental Studies.
- [13] Daniels, P. W. (1972). "Transport Changes Generated by Decentralized Offices." *Regional Studies*, 6: pp. 273–289.
- [14] Ewing, R. (1995). *Best Development Practices: Doing the Right Thing and Making Money at the Same Time*. Chicago: Planners Press.
- [15] Federal Transit Administration. (2000). *Transit Benefits 2000 Working Papers: A Public Choice Policy Analysis*. Washington, D.C.: Federal Transit Administration, Office of Policy Development.
- [16] Frank, L. D. (1994). The Impacts of Mixed Use and Density on The Utilization of Three Modes of Travel: The Single Occupant Vehicle, Transit, and Walking. Paper presented at the 73rd Annual Meeting of the Transportation Research Board, Washington, DC (January 9–13, 1994).
- [17] Frank, L. D. and Gary Pivo. (1994). *Relationship Between Land Use And Travel Behavior in the Puget Sound Region*. Olympia, WA: Washington State Department of Transportation, WA-RD 351.1.
- [18] Garrett, T. A. (2004). Light-Rail Transit in America: Policy Issues and Prospects for Economic Development. Federal Reserve Bank of St. Louis (<http://www.cfte.org/news/garrett.pdf>).
- [19] Giuliano, G. (1991). "Is Jobs-Housing Balance a Transportation Issue?" *Transportation Research Board Record 1305*. Washington, DC: Transportation Research Board.
- [20] Hooper, K. G. and JHK and Associates. (1989). *NCHRP Report 323, Travel Characteristics at Large-Scale Suburban Activity Centers*. Washington, DC, Transportation Research Board.
- [21] JHK and Associates. (1987). *Development-Related Ridership Survey I*. Washington, DC: Washington Metropolitan Area Transit Authority.
- [22] Landis, J., S. Guhathakurta, and M. Zhang. (1994). Capitalization of Transit Investments into Single-Family Home Prices: A Comparative Analysis of Five California Rail Transit Systems. Monograph 246. Berkeley: The University of California Transportation Center.
- [23] Landis, John, R. Cervero, S. Guhathakurta, D. Loutzenheiser, and M. Zhang. (1995). Rail Transit Investments, Real Estate Values, and Land Use Change: A Comparative Analysis of Five California Rail Transit Systems. Monograph 48, Institute of Urban and Regional Studies, University of California at Berkeley.
- [24] Levinson, D. and A. Kumar. (1994). "The Rational Locator: Why Travel Times Have Remained Stable." *Journal of the American Planning Association*, 60, 3 pp. 319–332.
- [25] Mills, E. S. (1972). *Urban Economics*. Glenview, Ill.: Scott, Foresman and Company.
- [26] Muth, R. E. (1969). *Cities and Housing: The Spatial Pattern of Urban Residential Land Use*. Chicago: The University of Chicago Press.
- [27] Nelson, A. (1992). Effects of Elevated Heavy-Rail Transit Stations on House Prices with Respect to Neighborhood Income. *Transportation Research Record* 1359, pp. 127-132.
- [28] Nelson/Nygaard Consulting Associates. (1995). "Land use and Transit Demand: The Transit Orientation Index." In Chapter 3 of *Primary Transit Network Study*. Portland, OR: Tri-Met
- [29] Nowlan, D. and G. Stewart. (1991). "Downtown Population Growth and Commuting Trips: Recent Experience in Toronto." *Journal of the American Planning Association*. 65 pp. 165–182.
- [30] Rice Center for Urban Mobility Research. (1987). *Assessment of Changes in Property Values in Transit Areas*. Houston, Rice Center.
- [31] Transit Cooperative Research Program. (1996). *Transit, Urban Form, and the Built Environment: A Summary of Knowledge*. In *Transit and Urban Form Volume 1*. Transportation Research Board, National Research Council. Washington D.C.: National Academies Press.
- [32] Wachs, M., B. Taylor, N. Levine, and P. Ong. (1993). "The Changing Commute: A Case-study of the Jobs-Housing Relationship Over Time." *Urban Studies*, 30, 10 pp. 1711–1729.
- [33] Waddell Paul and Borning Alan. 2004. [A Case Study in Digital Government: Developing and Applying UrbanSim, a System for Simulating Urban Land Use, Transportation, and Environmental Impacts](#), *Social Science Computer Review*, 22 (1): 37-51. <http://www.urbansim.org/papers/index.shtml>.
- [34] Waddell Paul and Ulfarsson Gudmundur F. 2004. Introduction to Urban Simulation: Design and Development of Operational Models. Forthcoming in Handbook in Transport, Volume 5: Transport Geography and Spatial Systems, Stopher, Button, Kingsley, Hensher eds. Pergamon Press. www.urbansim.org/papers
- [35] Waddell, Paul. (2002). [UrbanSim: Modeling Urban Development for Land Use, Transportation and Environmental Planning](#). *Journal of the American Planning Association* 68 (3): 297-314.
- [36] Weinstein, B. and T. Clower. (1999). *TCRP Legal Research Digest 12: The Initial Economic Impacts of the DART LRT System*. Washington, D.C.: Transportation Research Board of the National Academies.

Intelligent Cities

Steve Curwell

University of Salford

Research Institute for the Built and Human Environment

4th Floor, Maxwell Building

Salford, M5 4WT, UK

+44 161 295 4622

s.r.curwell@salford.ac.uk

1. INTRODUCTION

This paper will report on the main outcomes of the INTELCITIES Intelligent Cities integrated project (IST no. 507860), which commenced in January 2004. The “vital statistics” of the project can be summarised as:

- Budget: Euro 11.7M (EU Contribution 6.8M);
- Based on the INTELCITY FP5 Roadmap project exploring how the knowledge society can be achieve by 2010 and sustainable urban development by 2030 - as reported in eChallenges 2004;
- Critical mass of 18 cities, 20 ICT companies, 36 research groups including 16 SMEs in a total of 20 European Countries;
- Prototype e-Government service modules being “built” in a number of European cities including Marseille, Siena, Helsinki, Rome, Leicester, Dresden, Berlin and Manchester. The modules are being linked together to demonstrate an Integrated Open System eCity Platform (e-CP);
- Methodology based on iterative learning where R & D pilot studies are embedded in cities and are meeting citizens’ needs;
- Recognition of the need for new business models, e.g. PPPs which offer new ways of delivering services and business opportunities;
- Coordinator – City of Manchester: Dave Carter, Head of Digital Development Agency and Scientific & Technical Direction - University of Salford, UK: Prof Steve Curwell.

2. OBJECTIVES

The main objective of INTELCITIES is to create a new and innovative set of interoperable local e-government services to meet the needs of both citizens and businesses through the development of an e-City Platform. This will provide interactive citywide on-line applications and services for users that will make all aspects of what is “going-on” in the city available to all. This

will support:

1. the everyday needs and requirements of citizens and business through an 24 hour access to enhanced transactional city services;
2. more efficient city management and administration by integrating functions and services across city authorities, regional and national governmental agencies, utility and transport system providers and citizens/NGO networks;
3. much more innovative and effective approaches to urban planning through more reliable electronic city modelling, using advanced visualisation and predictive techniques, which will enable citizens and businesses to play a far more participative and inclusive role in influencing how planned changes in the city will affect their lives.
4. New pan-European e-government services addressing employment, business re/location, events management, civil defence and environmental protection

3. METHODOLOGY USED

New forms of electronic governance are at an experimental stage in which learning by doing in cities is the key, especially in terms of developing and utilizing best practice. In the project prototype system development has been undertaken in cities in “living lab” test bed conditions.. The project is showing how various transactional electronic services and technology platforms, e.g. iDTV, PC based and mobile, in diverse socio-economic environments applications and services can be integrated and made interoperable in ways that provide business intelligence. This involves both technological advance in the development of the e-city platform middleware as well as addressing aspects of back-office re-organisation in cities and capacity building in civil servants, businesses and citizens to extend social inclusion. Experiments in both mobile and wired technologies are demonstrating how enhanced services can support development of new local government e-services and enable citizens to participate more fully in the information society and the knowledge economy.

4. TECHNOLOGY CASE DESCRIPTION

Integration and interoperability form the main technical challenges in the project. The modular structure used to develop a number of e-services for both day-to-day city management and medium to long term city planning will be explored. The paper will describe the three level technical architecture used for the middleware consisting of and as shown in figure 1:

e-City Platform conceptual architecture

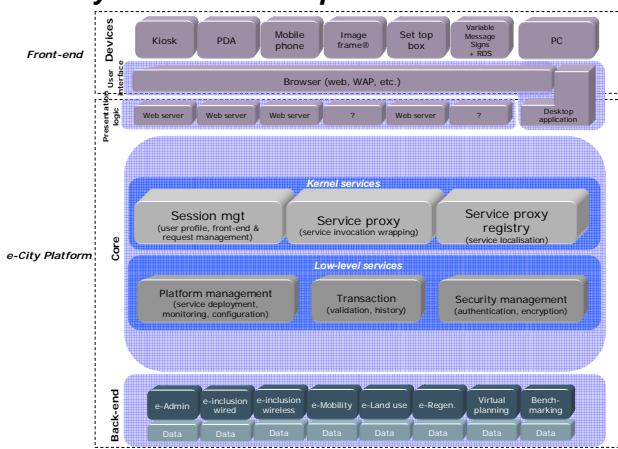


Figure 1. System Architecture
<http://www.intlcitiesproject.com>

- Front-end – including provision for all interface devices required by users including PC, PDA, iDTV, Kiosks, etc. .
- Core middleware – including the kernel services of session management, service proxy (wrapping) and registry with the low level platform management, transaction and security management;
- Back-end – consisting of all the e-government service modules developed in the project embracing e-administration, citizen participation and voting, mobility, GIS, planning and urban regeneration.

Intlcities is unique in that it allows services to be spatially located, e.g. it is possible not only for a job seeker to automatically ascertain the location of the job, but also to access data public transport and on housing provision and its location.

5. CONCLUSIONS AND SUMMARY RECOMMENDATIONS

The key outcomes will be described in terms of:

- user needs and requirements studies undertaken in participating cities;
- prototype studies for a range of new or enhanced transactional services that have been designed, including software and hardware and the developing know-how emerging from experimental deployment in the host cities
- the e-City Platform middleware and its use to integrated local e-Government services and the potential for the business intelligence that is emerging.
- an interactive assessment tool, containing indicators allowing cities to benchmark their progress on e-governance, knowledge society and sustainable development
- a knowledge management toolkit containing the good practice or “orgware” developed in the project to enable cities to deploy the e-City platform, including e-learning, capacity development and back office reorganization.

6. FUTURE DEVELOPMENTS

The outcomes of the project in terms of the middleware solution and the various e-Government and -ePlanning services are to be marketed through an umbrella organization – the Intlcities Alliance, which will be described. The research findings are to be maintained through the development of a new European research centre on Intelligent Cities. This is to be a virtual centre with a secretariat based in the city of Siena, Italy.

SESSION 5A

TECHNOLOGY TRANSFER

Moderator

Rene Wagenaar, Delft University of Technology, The Netherlands

Titles and Authors

eGOVERNMENT - An European eGovernment Research Funding Agency Network
Knudsen, Trond; Siösteen-Thiel, Madeleine

Multidisciplinary E-Government Research and Education as a Catalyst for Effective Information Technology Transfer to Regional Governments
Vélez-Rivera, Bienvenido; Fernández-Sein, Rafael; Rodríguez-Martínez, Manuel; Rivera-Vega, Pedro I.; Díaz, Walter; Nuñez-Molina, Mario

Challenges in eGovernment Technology Transfer
Chun, Soon Ae; Yesha, Yelena; Adam, Nabil; Atluri, Vijay

COSPA (Consortium for studying, evaluating, and supporting the introduction of Open Source software and Open Data Standards in the Public Administration)
Rossi, Bruno; Russo, Barbara; Succi Giancarlo

eGOVERNET – An European eGovernment Research Funding Agency Network

Madeleine Siösteen-Thiel

eServices in Public Sector

VINNOVA – Swedish Agency for Innovation Systems
SE-10158 Stockholm, Sweden
Tel. +46 8 4733142

Madeleine.Siosteen-Thiel@VINNOVA.se

Trond Knudsen

R&D for iGovernment, Research Council of Norway
P.o.Box 2700 St. Hanshaugen
NO-0779 Oslo, Norway
Tel. +47 91703728
tk@rcn.no

ABSTRACT

“eGOVERNET – The European eGovernment Research Network” [1] is an EU financed ‘coordination action’ project for mapping eGovernment R&D financing initiatives across the EU and other European nations. The project will also develop an overview of eGovernment R&D in Europe, develop insights and networks across interest groups, suggest programme co-ordination tasks, and develop a framework for eGovernment R&D work. The paper highlights the project beginning phases and the work done to date. The ambition of building an eGovernment network and creating an overview of relevant research programmes, instruments and funding possibilities open outside Europe also is addressed.

Categories and Subject Descriptors

[Digital Government]: Research coordination and network building – *mapping of governmental research funding, programmes and instruments.*

General Terms

Management, Documentation, Legal Aspects, Verification.

Keywords

Digital government, eGovernment. Research co-operation and co-ordination.

1. INTRODUCTION

The project “eGovernet – The European eGovernment Network” is financed by the EU 6th Framework Programme within Information Society Technologies as a coordination action starting January 2006 and running for two years. Ministries and agencies financing eGovernment research from seven European nations together with the EU’s Institute for Prospective Technological Studies – IPTS, in Seville, Spain constitutes the project’s partners. Sweden’s VINNOVA is coordinating the work.

An Interest Group of active non-partners is under construction with actors and key players from public administrations, research, and industry. This work is ongoing and will support the activities of the core partners.

2. BACKGROUND

Fragmented research-funding mechanisms, low visibility and low, or absent, trans-national coordination of national and regional eGovernment-research policies are some of the obstacles on the route to establishing a European Research Area (ERA) in the

eGovernment domain. European eGovernment research clearly can benefit from improved support by a research strategy on the European level.

Table 1. The Project Partners

Institution	Representing
Academy of Sciences of the Czech Republic - ASCR	Czech Republic
Information Society Development Committe Govt Lithuania - ISDC	Lithuania
Information Society Policy Division, Department of the Taoiseach, - DEPTAO	Ireland
Institute for Prospective Technological Studies, - IPTS	EU
Research Council of Norway - RCN	Norway
Slovenian Ministry of Higher Education - SLOMIS	Slovenia
The Ministry of Science and Information Society Technologies, - MNII	Poland
VINNOVA - Swedish Agency for Innovation Systems	Sweden

The main objective of the eGOVERNET project is to coordinate the creation of national eGovernment RTD programmes/initiatives as well as stimulate the integration of existing national eGovernment programmes. The consortium behind the eGOVERNET project represents organisations with national programme responsibilities for eGovernment innovation and research in their respective countries, including both old and the new member states and associated states. The partners have come together with a vision to coordinate their research policies and work towards a long-term strategy for eGovernment research.

The results of the project will include studies of state-of-the-art, best practice in eGovernment, and impact indicators, together with a comprehensive knowledge resource directory. This will feed into a framework for a common European eGovernment research agenda.

The eGOVERNET project will work towards the development of a shared framework for eGovernment RTD targeting coordinated RTD activities, policies and thematic foci, to promote innovation by synergy in European eGovernment Research.

3. PROJECT RATIONALE AND OVERALL OBJECTIVES

The eGOVERNMENT consortium is formed from an existing network of organizations, which was informally established in 2004, in order to exchange ideas and requirements with respect to the needs for improved cooperation and resource allocation for eGovernment related research. The consortium behind the eGOVERNMENT project includes organisations that are either directly responsible for national RTD funding, or, are responsible at a strategic level for defining or implementing policies regarding eGovernment. The consortium includes both the old and the new member states and an associated state.

During the first year of cooperation it became clear that eGovernment research in Europe suffers from a lack of visibility, has fragmented funding mechanisms and lacks coordination of national research policies. The different member states are at different stages in developing eGovernment research initiatives. A few countries have specific eGovernment research programmes, while in other countries there is no established research area, but funded as part of other programmes. Other member states have initiatives to create national eGovernment research programmes on the way. This varying level of support and research mass is a major obstacle to the establishment of a European Research Area for eGovernment.

Thus, the organisations recognised a need to exchange experiences, harmonize views, discuss research priorities, compare best practices, and establish benchmarking principles and impact indicators. As a result the consortium proposed a coordination action to work towards establishing a long-term strategy for innovative eGovernment research.

The main objective of the eGOVERNMENT project is to coordinate the creation of national eGovernment RTD programmes/initiatives as well as stimulate the integration of existing national eGovernment programmes. The eGOVERNMENT coordination action is establishing links to ongoing and future EU-funded research projects in the area.

Pursuing the goals of the Lisbon agenda and the i2010 vision, requires an improved coordination of research initiatives to promote innovation in research programmes, to leverage the involvement of stakeholders (public sector, research, industry), and to realize a greater use of information services such as PEGS (pan-European eGovernment services). At mid-term in the implementation of the Lisbon agenda, there is a lack of feedback to the research community on these developments, and the rate at which research results are being applied, is too slow. Although there are many eGovernment initiatives in Europe with significant investment, there is a need

- to increase involvement by bringing researchers, industry, end-users and public sector closer together
- to stimulate innovation in national research programs
- to explore opportunities for opening windows to eGovernment for other research areas

- for a structured and common view of objectives, results and visions for future research

The eGOVERNMENT project involves many different levels of cooperation and coordination. The results of the project will include studies of state-of-the-art and best practice in eGovernment together with a comprehensive knowledge resource directory. This will feed into building a framework for a common research agenda for eGovernment in Europe, as basis for future research strategies.

eGOVERNMENT provides an arena and the sustainable instruments for increased eGovernment research cooperation. The eGOVERNMENT project will

- facilitate the development of an RTD framework for eGovernment by involving key stakeholders such as RTD funding and policy making organisations.
- assist in the development of a focal area of European eGovernment research and innovation.

The overall objective of the eGOVERNMENT coordination action is to intensify the coordination of eGovernment research activities carried out at national or regional level in the Member States and Associated States and to improve co-operation among the European eGovernment research community through:

- organized mutual information exchange on national and regional research policies, programmes and thematic foci.
- suggestions for networking of research activities conducted at national or regional level, with those on the EU-level.

The project adopts a long-term perspective in its work, appreciating the fact that the organization of research varies over time and place and in order to ensure that results are durable beyond the project period. eGovernment is a relatively new area for research. There are few existing national programmes focusing specifically on eGovernment. Whereas research related to eGovernment certainly exists, it is scattered over different existing research programmes, and we lack the big picture relating this research to the needs of the national and EU-levels.

At this time, re-starting the Lisbon agenda with new and fresh visions, there is a window of opportunity for an effective coordination action, with a possibility to promote the instigation of new national initiatives and programmes, for eGovernment.

4. ACKNOWLEDGMENTS

Our thanks to all eGOVERNMENT partner representatives that has contributed to the project work, from which this highlights are deducted.

5. REFERENCES

- [1] Sjösteen-Thiel, M. et al.: eGOVERNMENT – The European Network of eGovernment, Project Description. VINNOVA Preliminary Report (Nov. 2005).

Multidisciplinary E-Government Research and Education as a Catalyst for Effective Information Technology Transfer to Regional Governments

Bienvenido Vélez-Rivera
Rafael Fernández-Sein

Manuel Rodríguez-Martínez
Pedro I. Rivera-Vega

Walter Díaz
Mario Núñez-Molina

Department of Electrical and Computer Engineering
University of Puerto Rico Mayagüez
egov@ece.uprm.edu

ABSTRACT

The Digital Government project at the University of Puerto Rico Mayagüez consists of research and education components aimed at identifying the most pressing obstacles to the adoption of information technologies by small regional governments in Puerto Rico. The project will propose optimal technological paths designed to ameliorate if not eradicate such obstacles. The end goal is to help regional governments adopt software technology suites that are reliable, accessible, cost effective and capable of empowering citizens with closer and better governmental services.

Categories and Subject Descriptors

K.4 [Computers and Society]: Digital Government

General Terms

Design, Experimentation, Security, Human Factors, Legal Aspects.

Keywords

Digital Government, Heterogeneous Databases, Multi-lingual, Information Retrieval, Hierarchical XML schemas.

1. INTRODUCTION

Electronic government systems have an unprecedented potential to improve the responsiveness of governments to the needs of the people that they are designed to serve. To this day, this potential is barely beginning to be exploited. Significant barriers hinder the effective integration of information technologies into government practices and their adoption by the public. Government agencies often find themselves in a disadvantaged position to compete with the private sector for information technology workers, a workforce

whose shortage at a national level is well recognized. The need to abide by rigid procurement practices makes it virtually impossible for agencies to keep their technology infrastructure up to date with the fast pace of technological advances. For instance, Amdahl's law, a well known technological trend, predicts that processor speed doubles approximately every 18 months. Local and regional governments are particularly affected by this state of affairs.

The Digital Government Research project at the University of Puerto Rico Mayagüez is an NSF-funded effort by a multidisciplinary group including researchers from the University of Puerto Rico-Mayagüez (UPRM) and personnel from the municipal government from the city of Mayagüez. The group members combine their talents in Public Administration, Computer Science, Engineering and Social Sciences, in order to: identify significant barriers to the effective transfer of information technology into government practices and their adoption by the public, engineer novel solutions to help overcome these barriers, and test their solutions in a real municipal government environment.

2. Technology Transfer Component

This component consist of a series of annual software projects in which teams of senior undergraduate students under the direct supervision of a faculty member get involved with an agency from the city of Mayagüez, the UPRM's hosting town, in the design of web-enabled applications aimed at satisfying the highest priority information needs of this agency. Each team of students goes through all the phases of the software engineering project from requirements elicitation, to design, testing and finally deployment.

During the third year of project we continued improving the applications for the three major Mayagüez agencies: Public Works, Citizens Services and Public Housing. We have also begun the design of a new application allowing people to conduct transactions involving payments for Municipal services, permits, and taxes among others. We expect a first prototype of the system to be available in April of 2006.

3. RESEARCH COMPONENT

3.1 Multi-lingual Document Repositories

Our group is developing an information retrieval engine supporting and XML-based query language novel in two ways. First, the engine allows the expression of queries based on virtual hierarchical XML schemas encompassing several similar concrete schemas. Current systems force the user to specify one XML query for each schema available in the database thus limiting query writeability to the a priori knowledge that the user has of the available XML schemas. Second, our system will support the dynamic integration of new XML schemas as these are developed and made available to the public. This adaptation to new schemas will no require costly modification and recompilation of the search engine thus yielding a more available and easier to maintain system.

3.2 Heterogeneous Governmental Databases

During the second year the effort on heterogeneous databases has focused on the design and development of the NetTraveler system. NetTraveler is a middleware solution supporting the orchestration, choreography and composition of web services to assemble applications from pre-deployed components. The effort pawns research threads in peer-to-peer database-backed applications as well as in distributed query optimization.

3.3 Semantic Document Management

This component of our Digital Government project works in collaboration with the regional office of the Registry of Deeds in Mayagüez. The effort is aimed demonstrating the feasibility of current technologies such as XML, XForms and the Business Process Language as building blocks upon which workflow applications can be build in a cost effective manner to automate man of the processes run at the mentioned office. We are currently working on improving BPL's support for business processes that require direct synchronous or asynchronous interaction with human participants. In particular, we are interested in designing and developing plug-in's for commonly used email systems in order to allow government employees to interact with other automated workflow engine components using familiar interfaces and without requiring re-learning of new applications.

4. Annual Digital Government Congress

The Digital Government Project at the UPRM organized and celebrated the second annual Digital Government congress in Puerto Rico during the month of May of 2005. This second edition of the congress lasted two days and included a morning of presentations of both the technology transfer and the research components of the projects. Students presented their results in front of an audience which included people from the UPRM academic community, government personnel from several municipalities and state-wide governmental agencies. In one afternoon, UPRM Faculty organized free workshops tuned to the technology needs of the City of Mayagüez personnel as surveyed by Walter Díaz and Mario Núñez. The second day was spent offering several conferences by invited guest speakers and an afternoon panel on Digital Government challenges in Puerto Rico.

5. Accomplishments for Funding Year 2

- IT Systems for Mayaguez 100% operational in two municipal agencies: Public works and Citizens Services
- Phidelix Technologies Inc founded by project PI and CoPI's to offer long term support for IT systems.
- Collaboration started with PR Court Administration Agency
- Completely redesigned user interface for IT Systems
- Published research results in peer reviewed conferences

6. Challenges for Funding Year 3

- Complete deployment of Housing Office System
- Build first prototype of DG payments system
- Reach long term IT outsourcing commercial agreement with the City of Mayagüez
- Offer citizens direct online access to DG services
- Launch MiPuebloDigital.com portal as the cyber space of choice for municipal information and DG services in Puerto Rico

7. ACKNOWLEDGMENTS

We want to thank the City of Mayagüez and its Major José Guillermo Rodriguez for their support during the first two years of the project. We also want to thank the Registry of Deeds from the City of Mayagüez for providing us access to their information and knowledge. The Puerto Rico Bar Association has kindly agreed to collaborate with us on the analysis and design of standards for legal documents used by the Registry of Deeds as well as other governmental agencies.

8. REFERENCES

- [1] Caituiro, Hillary & Rodriguez, Manuel. *Net Traveler: A Framework for Autonomic Web Services Collaboration, Orchestration and Choreography in E-Government Information Systems*. IEEE International Conference on Web Services (ICWS 2004). San Diego, California, USA. July 6-9, 2004.
- [2] Rodriguez, Manuel, *Smart Mirrors: Peer-to-Peer Web Services for Publishing Electronic Document*. Proceedings of the 14th International Workshop on Research Issues on Data Engineering: Web Services for E-Commerce and E-Government Applications. Boston, USA , March 28-29, 2004.
- [3] Velez, Ivan & Velez, Bienvenido. *Lynx: An Open Email Extension for Workflow Systems Based on Web Services and its Application to Digital Government* Proceedings of IEEE International Conference on Internet and Web Applications and Services (ICIW'06). Guadeloupe, French Caribbean. February 23-25, 2006.
- [4] Systems Velez, Ivan & Velez, Bienvenido *Lynx: An Open Architecture for Catalyzing the Deployment of Interactive Digital Government Workflow-Based Applications*. Submitted to 7th Annual International Conference on Digital Government Research (DG.O 2006). San Diego, California, USA. May 21-24, 2006.

Challenges in eGovernment Technology Transfer

Soon Ae Chun

City University of New York
Staten Island, NY
718-982-2931

chun@mail.csi.cuny.edu

Yelena Yesha

University of Maryland
Baltimore County, MD
410-455-3542

yeyesh@cs.umbc.edu

Nabil Adam, Vijay Atluri

Rutgers University
Newark, NJ
973-353-1014

{adam,atluri}@cimic.rutgers.edu

1. INTRODUCTION

Digital government has four distinct phases in its evolution [6]. In the first *Catalog* phase, governments publish information to view and download forms and other static documentation. The second *Transactional* phase facilitates information exchange and online payments. The third *Vertical Integration* phase provides integrated services within a governmental department. Finally, *Horizontal Integration* phase provides integrated services across various levels of governments and agencies: local, county, and state.

In another development, digital government has been shifting from a technology-driven tool to a value-driven tool, which focuses on citizen and business-centered services. This end-user-centric design entails the reduction of burdens on the user, including a reduction of the cognitive load, trying to minimize information overload, as well as time and effort for consuming the services. Provider-centered eGovernment primarily focuses on optimizing the business operations within government and the reduction of costs involved. This “human-centered eGovernment” requires considering actual and prospective users’ needs, such as content, accessibility and navigational preferences such as structure and modes.

Our projects with the New Jersey State government agencies have developed human-centric services for business communities [1][3] and technology transfer [2]. In this paper, we present the past and on-going experiences and challenges of technology transfer from research prototypes to full-scale operational government services, emphasizing the gaps to be filled for a successful technology transfer for eGovernment.

2. HUMAN-CENTERED BUSINESS SERVICES

Our projects have focused on the inter-agency services for large and small business communities, such as a new business registration process, a vendor registration process, and a business land development process, that required the vertical and horizontal integrations and coordination among multiple agencies within the State as well as across multiple governments. For instance, the land development compliance process in coastal areas requires interagency coordination from Departments of Environmental Protection, the U.S. Army Corps of Engineers, and local municipalities as well as permits from the New Jersey

Meadowlands Commission (NJMC). Its prototype, Spatially Integrated Coastal Permitting System (SICOP), helps developers to identify the required permits for coastal wetland development and alteration projects according to property location, spatial characteristics and the type of project; it defines the permit application submission sequence, and visualizes a “road map” for the applicant to follow, with appropriate information on application forms and agency websites.

Another prototype provides the State government’s business-related services (business registration, vendor registration, licensing & permits services). It involves coordination by Departments such as Treasury, Commerce, Environmental Protection, Public Health and Safety, and Divisions such as Commercial Recording, Community Affairs, Unemployment and Disability Insurance, Worker’s Compensation, and many others.

These research prototypes for our human-centered approach have the following characteristics:

- *One-stop shop portal approach*: Business communities could have a single interface portal that facilitates locating the relevant information and services.
- *Simple and Intuitive interface*: The business owners and developers use a step-by-step question answering interface as well as visualization of information, as much as possible, through maps, workflow graphs and directory structures.
- *Data sharing to avoid data redundancies*: Data captured from the users are shared among multiple services and among multiple agencies.
- *Knowledge-based personalization*: An ontology is used to capture the regulatory knowledge required for business-related services and the government officials’ operational intelligence on how and where these regulations apply [3].

The goals of the projects have included (1) creation of value chains for the business communities; (2) operational efficiency enhancements in the governments; (3) automatic compliance checks against regulatory mandates; (4) promotion of data sharing, breaking the “stovepipe” paradigm of the government processes.

The SICOP prototype system for business’s land development-related permit identification and permit application services has been successfully implemented as an operational service at New Jersey Meadowlands Commission [5]. The technology transfer of the State prototype for new business registration, vendor registration, and e-filing services are still in progress. In the following section, we discuss the success factors and challenges in the technology transfer.

3. TECHNOLOGY TRANSFER SUCCESS FACTORS AND CHALLENGES

The SICOP technology transfer has the following combination of success factors.

- Strong administration and management endorsement: The NJMC management level was able to convince permit-related agencies to be involved and support the implementation. Through meetings, the scope of the project was limited to what can be achieved and it was determined what can be delayed. The strategic decisions at the management level were made based on the practicality and reality of the political environment: which agencies are able to cooperate and which ones are not up to it.
- Users' and experts' involvement from the beginning to the end: The development was an iterative process where the end users and permit officials participated in every stage of the development with feedback on the interfaces and advice on permit information and processes. In the end, the technology transfer was one step in the development process.
- The cost of technology transfer was minimal: NJMC had full control of their computing resources, and an additional application did not require huge investment costs, or coordination costs.

The implementation of the State government project also follows a similar pattern.

- The management level (Chief Technology Officer, Treasurer) was behind the project, providing guidance and identifying the focus areas.
- We took into consideration the feedback from entrepreneurs visiting the NJ Small business centers.

However, during the technology transfer of the State government project, the government experts' involvement was comparatively small, considering the number of agencies involved. The expertise is distributed over a much larger number of government agencies. The regulatory ontology development is a huge undertaking that requires precise definitions and operational knowledge. The loose working relationship between government experts and developers may have to be changed to much tighter working relationships in order to make much smoother technology transfer.

Each agency had an operational streamlining within its organization, and it was found to be quite difficult to change their work habits. They all acknowledge the need for and benefits of the change, especially the benefits from coordinating with other agencies. However, it became difficult for them to conform to the changes. For instance, the business accounting process between the Motor Vehicle Commission and the Division of Revenue required document exchanges in order for preparing the annual summary reports. The usual data sharing in a PDF report format needed to change to the electronic raw data. This involved special actions by the data-managing center of another agency. It was proven difficult to provide the data in the requested format, slowing down the technology transfer task.

Another big challenge was the data ownership by each agency. This sense of data ownership prevented the academic developers from accessing and processing data for implementing cross-agency processes. An intermediate data format had to be prepared for the technology transfer stage.

In addition, the following challenges need to be addressed for an effective technology transfer:

- The governments collect and maintain huge amounts of private records of personal information of citizens (as in E-file tax

data), and they are as accountable for their confidentiality as any corporate businesses. The trust in the government by the citizens is based on their capability to prevent unauthorized access, to prevent invalid transactions from occurring and to maintain the integrity of the system.

- The partnership between the government agencies and the development team should be based on a strong formal agreement with more accountability as with the private business partners.
- Another challenge is that data sharing practices are quite informal between agencies, some based on long time practices, some more ad-hoc, and some non-existent. The data sharing policies in the State agencies may need to be codified such that the data sharing should be mandated and clearly defined. These sharing policies may overcome the ownership perception of the data by each agency.
- The governments operate with limited resources, but the technology transfer from academic partners will require dedicated government human resources to tightly work with. The hardware and software resources were found to be available at the State level, but are not easily accessible by one agency. As opposed to the NJMC case, each State agency resorted to a central technology agency and did not have full control of the computing resources, thus it required additional communication and coordination for the technology transfer.

4. CONCLUSION

In conclusion, the value chain processes can be beneficial for the citizens and users, but these user-centric one-stop processes impose operational integration across multiple agencies and require tighter working relationships and trust among partners for successful technology transfer. The challenges over data sharing and resource availabilities in timely manner are crucial for successful technology transfer.

5. ACKNOWLEDGMENTS

This work is partly supported by the National Science Foundation under grant IIS 0443591, and partly by the NOAA Coastal Service Center Grant program (Award Number NA17OC2587).

6. REFERENCES

- [1] N. Adam et. al. eGovernment: Human Centered Systems for Business Services *Proceedings of the First National Conference on Digital Government*, Los Angeles, CA, 2001.
- [2] N. Adam, V. Atluri, S. Chun, P. Fariselli, J. Culver-Hopper, O. Bojic, R. Stewart, J. Fruscione and N. Mannocchio, Technology Transfer of Inter-Agency Government Services and their Transnational Feasibility Studies, *Proceedings of the NSF dg.o 2005 Conference*, Atlanta, Georgia, May 15-18, 2005.
- [3] N. Adam, F. Artigas, V. Atluri, I. Bora, R. Ceberio, S. Chun and A. Paliwal, Spatially Integrated Coastal Permitting System (SICOP), in *Proceedings of the Coastal Zone 03 (CZ 03)*, Baltimore, Maryland, July 2003.
- [4] S. Chun, V. Atluri and N. Adam, Domain Knowledge-based Automatic Workflow Generation, in *Proceedings DEXA 2002, Lecture Notes in Computer Science 2453*, 2002.
- [5] Meadowlands Coastal Permitting Assistant: <http://meri.njmeadowlands.gov:8080/sicop/>
- [6] Donald Norris and M. Jae Moon. Advancing E-Government at the Grassroots: Tortoise or Hare? University of Maryland, Baltimore County. 2003.

COSPA (Consortium for studying, evaluating, and supporting the introduction of Open Source software and Open Data Standards in the Public Administration)

Bruno Rossi

Free University of Bolzano-Bozen
Piazza Domenicani, 3
39100 Bolzano-Bozen, Italy
Bruno.Rossi@unibz.it

Barbara Russo

Free University of Bolzano-Bozen
Piazza Domenicani, 3
39100 Bolzano-Bozen, Italy
Barbara.Russo@unibz.it

Giancarlo Succi

Free University of Bolzano-Bozen
Piazza Domenicani, 3
39100 Bolzano-Bozen, Italy
Giancarlo.Succi@unibz.it

ABSTRACT

In this paper, we report about COSPA, the Consortium for studying, evaluating, and supporting the introduction of Open Source Software (OSS) and Open Data Standards (ODS) in the Public Administration. The project, an EU FP6 research project, aims to evaluate the effects of the introduction of OSS and ODS for personal productivity and document management in European PAs. The objectives of the project and the major research activities will be reported, with particular relevance given to the current challenges and future research.

Keywords

OSS, Open Source Software, Migration, COSPA.

1. INTRODUCTION

COSPA, the Consortium for studying, evaluating, and supporting the introduction of Open Source Software (OSS) and Open Data Standards (ODS) in the Public Administration, is an EU FP6 research project that aims to experimentally assess the potential benefits for European PAs that plan to introduce and evaluate Open Data Standards and Open Source Software in desktop applications for personal productivity within the context of European PAs.

2. CONSORTIUM DETAILS

The consortium is coordinated by the Free University of Bozen-Bolzano and is constituted by 4 Research Institutions (Aalborg University (Denmark), Sheffield University (UK), Limerick University(Ireland), MTA Computer and Automation Research Institute (Hungary), plus 8 partner PAs, 2 partners from the industry (Conecta, IBM) and more than 60 international observers with the aim to disseminate the results of the consortium. More than 19 European countries are represented inside the consortium.

2.1 Consortium objectives

The aims of the consortium are related to the study and introduction of OSS and ODS inside the PA, in particular the objectives are the following:

- to analyze the requirements of PAs in order to ease the deployment of OSS. In this sense the objective of the consortium is to identify possible OSS and ODS solutions that fulfill the gathered requirements.
- to perform pilot installations and experimentations with OSS on the desktop side of partner PAs, to ease the successive cost/benefit analysis.

- to evaluate OSS through a rigorous statistical and cost/benefit analysis. The focus is to determine whether OSS can be a viable solution to adhere to the requirements
- to build a European knowledge and experience repository useful for migration purposes.
- to disseminate the results and the experiences of the project through the knowledge base and a series of workshops at regional and European level, with the aim of stimulate the sharing of knowledge and the public and business' awareness on the project and on OSS in general.

3. RESEARCH ACTIVITIES

The interactions between the different parts of the project are reported in Figure 1.

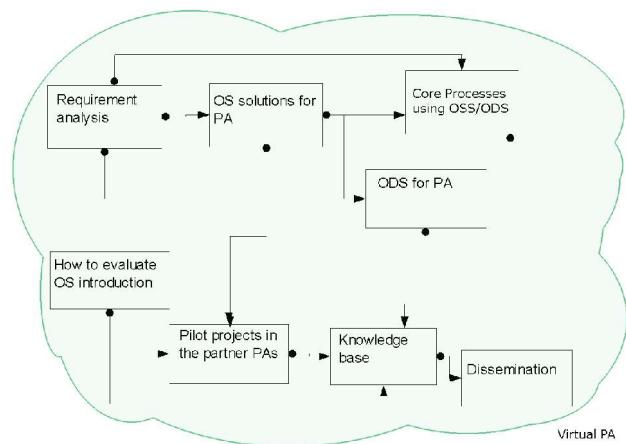


Figure 1. Interactions between different parts of the project.

The project initial research activities performed regarded the extraction of the requirements from the partners PAs and the identification of the possible solutions for the deployment of OSS and ODS. This activity resulted in one large catalogue of OSS and one on ODS that will be published as soon as the deliverable receives the Commission final approbation. A large work of the consortium was related to the acquisition of the requirements at the business process (BP) level, BPs were extracted from partner PAs to determine the exact needs for the activities. All the

subsequent analysis has been and will be performed at this level of detail. A framework is currently under development to evaluate the migration under a cost and benefit analysis and the quantitative data coming from the pilot introductions of OSS and ODS in the partner PAs will be inserted into the framework to provide a real evaluation of the migration process. The deployment of the experimentation has been carried on over 4,500 sites with a monitored transition to the new desktop solution; this large scale investigation will give the consortium enough data to evaluate the migration. All the activities are complementary to the creation of a European knowledge base on OSS that should serve citizens and organizations willing to acquire information on open solutions and on real case studies of migrations.

A characteristic of the analysis so far, is that to facilitate the understanding of a PA's roles and actions, the concept of Virtual PA has been created. It is a synthetic model of a PA refined from all the real-world data collected during the project. This conceptual PA has proved useful for simplifying the process of categorizing needs and resources into a manageable model.

4.FUTURE WORK

While the results are so far encouraging, many are the issues still open in the project. Such a large deployment of OSS for studying purposes has never been done and requires a rigorous framework for the successive analysis. The choice has been to model the framework according to the TCO methodology.

4.1Successes and failures so far

The collection of the requirements in particular at the Business Process level has been a successful activity so far. One constant problem of the consortium has been to obtain the results under the given schedule. Some activities, like the collection of the requirements and the running of the pilot introduction of OSS/ODS required more time as was foreseen, also due to the large amount of data to be collected and of the different partners involved. Differences in culture, distance and also bureaucracy all played a major role in the final delays.

The dissemination results were positive, in particular the observer program convinced many companies across Europe to join the project contributing to diffuse further information about the project, OSS and ODS.

4.2Challenges

It remains to be seen whether the project can meet all the initial aims that were posed by the European Commission. The next months will be crucial for finalization of the deliverables and the results will be available to the public domain.

4.3Plans to the end of the project

The next months will be of key importance for two parts of the project in particular: the finalization of the framework for the evaluation of the migration and the deployment of the COSPA knowledge base. All the parts of the project will merge together in order to produce the final deliverables. To note that the original point of the project was not to probe the superiority of OSS over

Closed Source Software or vice versa, but to assess whether OSS can be seen as a valid alternative to current software installations in PAs across Europe.

5.PAPERS PUBLISHED

Several papers have been submitted to European and International conferences, journals and magazines:

[1] An Experience of Transition to Open Source Software in Local Authorities, by P. Zuliani and G.Succi. Accepted for e-Challenges 2004 Conference, 27-29 October 2004, Vienna, Austria (<http://www.echallenges.org>)

[2] Exploiting the Collaboration between Open Source Developers and Research, by G. Succi and P. Zuliani. Accepted for ICSE 4th Workshop on OSS Engineering, 25 May 2004, Edinburgh, UK (<http://opensource.ucc.ie/icse2004/>)

[3] La Migrazione delle Pubbliche amministrazioni verso l'open source, by P.Zuliani, B. Russo, A. Sillitti, G. Succi. Accepted for AICA 2004 Conference, 28-30 September 2004, Benevento, Italy (<http://www.aica04.unisannio.it/>)

[4] Migrating Public administrations to open source software, by P. Zuliani and G. Succi. Accepted for IADIS e-Society 2004 Conference, 16-19 July, Avila, Spain (<http://www.iadis.org/es2004/>)

[5] Software aperto nella Pubblica Amministrazione: il progetto COSPA, by P. Zuliani, B. Rossi, G. Succi. Accepted for SALPA 2004 Conference, 22-23 March 2004, Pisa, Italy (<http://www.salpa.pisa.it>)

[6] Open Source Software and Open Data Standards in Public Administration, by G. Kovacs, S. Drozdik, P. Zuliani and G. Succi. Accepted for ICCC'04 Conference, 30 August – 01 September 2004, Vienna, Austria (<http://www.vmars.tuwien.ac.at/iccc04/>)

[7] Open Source Software for the Public Administration, by G. Kovacs, S. Drozdik, P. Zuliani and G. Succi. Accepted for the 6th CSIT (Computer Science and Information Technologies) International Workshop (Ed. N. Yussupova et al., ISBN: 5 87691-023-6), plenary paper, Budapest, October 17-19, 2004, Vol. 1. pp. 1-8.

[8] Risk Assessment of an Open Source Migration Project, by S. Drozdik, G. L. Kovács, P.Z. Kochis. Accepted for the OSS2005 Conference, 11-15 July 2005, Genova,Italy (<http://oss2005.case.unibz.it>)

6.REFERENCES

- [1] Main COSPA website, www.cospa-project.org .
- [2] COSPA Project of the Month, Europe's Information Society, http://europa.eu.int/information_society/activities/egovernme nt_research/projects/projects_of_month/200504/index_en.htm
- [3] COSPA Knowledge Base, <http://pascal.case.unibz.it>

SESSION 5B

E-RULEMAKING 1

Moderator

Stuart Shulman, Drake University, USA

Titles and Authors

Multidimensional Text Analysis for eRulemaking
Kwon, Namhee; Shulman, Stuart W.; Hovy, Eduard

Automatically Labeling Hierarchical Clusters
Treeratpituk, Pucktada; Callan, Jamie

Using Natural Language Processing to Improve eRulemaking
Cardie, Claire; Farina, Cynthia; Bruce, Thomas; Wagner, Erica

Multidimensional Text Analysis for eRulemaking

Namhee Kwon

USC Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292
+1-310-822-1511
nkwon@isi.edu

Stuart W. Shulman

University of Pittsburgh
121 University Place, Suite 600
Pittsburgh, PA 15260
+1-412-624-3776

shulman@pitt.edu

Eduard Hovy

USC Information Sciences Institute
4676 Admiralty Way
Marina del Rey, 90292
+1-310-822-1511

hovy@isi.edu

ABSTRACT

To support rule-writers, we are developing techniques to automatically analyze large number of public comments on proposed regulations. A document is analyzed in various ways including argument structure, topics, and opinions. The individual results are integrated into a unified output. The experiments reported here were performed on comments submitted to the Environmental Protection Agency in response to their proposed rule for mercury regulation.

Categories and Subject Descriptors

I.2.7 [Natural Languages]: Text analysis

General Terms

Algorithms, Experimentation, Human Factors, Languages

Keywords

Regulations, Electronic Rulemaking, Federal Government, Text Annotation, Semantic Frame, Argument Structure

1. INTRODUCTION

Every year thousands of government personnel at over 150 federal agencies and sub-agencies collaborate with stakeholder interest groups, lobbyists, lawyers, and citizens to craft as many as 8,000 regulations. The various actors involved increasingly use information and communication technologies in what has become known as electronic rulemaking, or e-rulemaking [30].

One former Director of the Federal Register recently warned that using the Internet to increase public participation in rulemaking “has inspired a sort of rulemaking arms race” in which some groups using web-based services “are convinced their position is strengthened by taking ten salient points and masquerading them as thousands of unique thoughts from thousands of thoughtful taxpayers.”[13].

Neil Eisner noted in his dg.o2005 keynote address that electronic rulemaking is a research domain teeming with eager government partners facing mounting information management challenges. At NSF-funded workshops and in numerous focus groups and

interviews, personnel at diverse federal agencies with significant rule writing responsibilities have indicated there is a dearth of tools for sorting through the comments on which modifications to proposed rules are supposed to be based [31]. As a result, headway is starting to be made in the various information retrieval tasks related to the automated sorting of large-scale public comment datasets [23][38].

The stakes are significant. Federal agency-issued rules are arguably much more important than the many fewer laws passed each year by the U.S. Congress [9]. Officials in the Bush Administration’s Office of Management and Budget reported in 2005 that the 190 major rules promulgated during the last 24 years collectively add well over \$100 billion annually in regulatory costs to the U.S. economy [26].

In certain exceptionally controversial cases, such as the Environmental Protection Agency’s recent mercury rulemaking, organizations such as MoveOn.org (a virtual organization with no office boasting over 3 million e-mail savvy members) seized upon the open and easily accessible “notice and comment” process (mandated by the 1947 Administrative Procedure Act) to issue sometimes voluminous demands for more stringent regulations. The data used in the experiments reported here were samples drawn from a population of over 536,000 e-mail messages and an additional 4,264 comments (consisting of some “good” e-mails, web form, fax, and paper submissions) that were submitted to the EPA about the proposed mercury rule. As one legal scholar rethinking regulatory democracy noted, “though individual members of the public who write comments usually make unsophisticated statements, those messages tend to include, at their core, constructive insights relevant to agencies’ legal mandates.”[12].

Given the sheer volume of comments and the time pressure on rule-writers, we are building a system to perform various types of analysis on the comments in order to provide a rich and multidimensional preview of a body of comments, and in order to help the rule-writers manage their work efficiently. The final goal of this work is to provide collective analysis of all comments including the quantitative and distributional analysis for each opinion and subtopics, and help the rule-writers to select more informative documents by highlighting the main focus of each document.

This analysis system consists of several independent modules. Importantly, the first step is to find duplicated comments and reduce them to a single instance, and to find near-duplicate comments and extract their idiosyncratic contents. This work is being performed by our colleagues at CMU, Jamie Callan et al.

[38]. Ultimately, we intend to combine the (near-)duplicate detection and the functionalities described in this paper into a single user-interface tool, for use by rule-writers in their exploration and analysis of a large collection of comments.

In this paper, we provide multidimensional text analysis exploring various aspects of text, currently focusing on topic, structure, and opinion. We analyze the writer’s argument structure about the proposed regulation, which is composed of a main claim over the proposal and reasoning to make its claim acceptable or justified. More specifically, the main claim is classified into “in favor of” or “against” the regulation, and the topics and concerns in the argument are grouped into several interesting subtopic codes. The problem is split into four subtasks: subtopic code detection, argument structure analysis, opinion analysis, and semantic frame analysis. We define each problem as a classification task and apply a supervised machine learning method using various features.

The rest of the paper is organized as follows: Section 2 defines the sub-problems and approach, Section 3 shows our manual annotation experience, Section 4 describes our automatic system and evaluation, and Section 5 concludes.

2. Approach

To handle the huge amount of comments, each document is translated into a unified structure and then integrated into collective result. We define several individual tasks to find semantic information, and tackle each problem separately. First, we define subtopic codes and classify each information unit (sentence) into appropriate codes. Second, we find a hierarchical argument structure to give preview in different detail, and then the opinion over the regulation is detected and analyzed into assent or opposition. Next, we provide the semantic frame analysis where each frame slot is representing the semantic role in a sentence. The final step is to integrate the output of each step and provide new collective result. By combining the argument structure with the opinion analysis, we can detect the reasons of each opinion. The specific tasks are described in the following sections.

2.1 Subtopic Codes

To develop the main idea, many topics or other concerns are introduced in public comments. We categorize the information into several subtopic codes defined in Table 1. For each information unit, the appropriate codes are selected (where applicable, multiple subtopic codes are assigned). These subtopic codes were developed by a political scientist (Stuart Shulman) and sociologist (Stephen Zavestoski) working deductively on previous theory-driven empirical work with public comments submitted to federal agencies in controversial environmental rulemakings.

Table 1. Subtopic Codes

Code	Text Description
Economic	Invokes economic concepts, such as cost, burden, benefits, growth, markets, efficiency, consumers, competitiveness
Environment	Invokes environmental preservation, intrinsic values, nature
Government responsibility	Invokes the responsibility of government to protect the public interest, preserve the rule of law, create procedural equity, transparent, consistent, clear-cut standards
Health	Invokes concerns about human health
Legal	Cites a particular statute, legal proceeding, administrative rule, or court case
Policy	Calls for a particular policy, such as the maximum achievable control technology (MACT) or a cap and trade approach
Pollution	Invokes the concept of pollution, human-made threats, ecological harm
Science	Invokes science, scientists, a specific study or scientific finding
Technology	Invokes technology

2.2 Argument Structures

Our assumption about the commentary texts is that a document has an internal argument structure to make its claim acceptable or justified. There might be a question whether the text is a real “argument” from the point of logic [27], but at least the comment contains claims and other statements supporting the claim, even if it may not be logical or valid.

We define the hierarchical tree structure of information units, each is a smallest part of text containing the coherent content as a claim or reasoning of the argument. The information unit is mostly a sentence but possibly a clause or a phrase. The hierarchical structure is limited to three levels, and each unit is classified into *root*, *subroot*, and *leaf* depending on its role in the argument. *Root* is a main claim over the proposal or conclusion of the argument. Multiple *roots* are possible in a document when more than one different claims are found, while there is only one root in most documents. Since the document is a comment on the proposed rule, we choose the most proposal-related claim as *root*. Each *root* has links to supporting reasoning, which is interpreted as a set of trees, each with *subroot* and *leaves*. *Subroot* is a direct supporting reason of the *root* and *leaves* are more detailed descriptions or examples of the *subroot*. All *leaves* under the same *subroot* share one topic.

In addition, we specify the types of relations between *root* and *subroot*, and between *subroot* and *leaf*. We classify a relation into three categories: *support*, *oppose*, and *restate*. *Restate* denotes the repetitive or paraphrasing statement of the parent. For example, when there are two important *roots* in a document but both of them make the same claim, we set one as a *root* and the other as a child of the *root* with the relation of *restate*. The *support* or *oppose* link shows whether the child supports or opposes the parent. Most links are *support* because people usually use positive reasoning for their claim. However, some texts contain conflicting reasons, for example, when a person mentions the pros and cons of his claim, or gives opposite

example of his claim as in the following sentence: (T_1 is linked to T_2 as a child with the relation of *oppose*.)

Although [the acid rain program that this cap and trade proposal is based on has been successful] T_1 , [it is simply an ineffective and inappropriate way to deal with mercury emissions.] T_2

Figure 1 shows an example of argument structure where each information unit is specified as *root*, *subroot*, and *leaf*, and *leaves* are expressed as subtopic codes.



Figure 1. Example Argument Structure of a document: leaves are shown as subtopic codes.

There have been various approaches to discourse structure analysis. Mann and Thompson [24] defined tree structure of discourse units having the status of *nucleus* or *satellite* with the rhetorical relations such as circumstance, contrast, and sequence. Teufel and Moens [32] defined non-hierarchical structure of scientific articles, more like a segmentation tagged with rhetorical function (background, aim, contrast, etc.) Our argument structure is similar to the tree structure of *nucleus* and *satellite* in [24], but the structure is the abstraction of argument about the regulation, so that the *root* and *subroot* relations are defined more globally in text rather than nested correlations between each discourse unit. In other words, finding *root* and *subroot* is to extract the claim and main reason in the document, which is comparable to extracting the “thesis statement” in [6].

2.3 Opinions

The public comments in our domain express the opinion or stance over the proposed rule. Our goal is to classify the comments if they are in favor of, or against the proposal, which is similar to finding the polarity (positive or negative) in previous research. Much work has addressed the problem of analyzing opinions from texts, including detecting subjectivity [35][39], classifying semantic orientation (polarity) of words [32], phrases [37], sentences [19][39], or documents [28]. Most approaches are based on lexical subjectivity and find dominant (more frequent or stronger) positive or negative expressions.

However, finding all positive and negative expressions is not enough. The example text (A) in Table 2 shows supportive (positive) expressions but actually opposes the rule because it compliments the alternative of the rule in question. Recognizing such cases requires topic analysis in addition to polarity analysis. Often, the topic is not available locally (expressed in a short phrase); instead, the specific contents of the rule or the alternative are described.

Further, one document possibly covers multiple topics and multiple opinions over the specific issues in the regulation (see the example (B) in Table 2). We attempt to detect every different opinion and to find corresponding reasons for each. Given the argument structure described in Section 2.2, we can assume that the main claim (opinion) about the proposed rule is in *root*. When there are different opinions in a document, each would be a separate *root*, and we determine the polarity of the *root* instead of whole text.

The opinion is classified into positive or negative attitude to the regulation. However, some texts suggest an alternative to the proposal or outside issues as in the example (C) in Table 2, so we define the opinion into three categories: *support the regulation*, *oppose the regulation*, or *propose a new idea*. The rule-makers may want to refer to the original full text when it is classified to *propose a new idea*.

Table 2. Opinion from Comments

- | | |
|-----|---|
| (A) | I support the recommendation by EPA staff scientists that both long- and short-term standards for fine particles need to be strengthened because scientific studies show serious health effects -- even death -- can occur at concentrations below the current standards. |
| (B) | The previous use of cap and trade methods for SOX removal is a good idea, and will work under the new mercury regulations. <u>I support the cap and trade method</u> , even if it may produce hot spots where more mercury settles...
The current plan proposed by President Bush lacks toughness when preventing mercury to be emitted into our air. The regulations set to 30% and 70% reduction by 2010 and 2018, respectively, are <u>too lenient</u> on power plants. Since the maximum available control technology can reach 70 to 90% mercury removal from stack gases if the removal is done efficiently, the regulations should be stricter. |
| (C) | We request that you extend the comment period either until June 30, or until 30 days after the completion and public availability of any new analysis, whichever is later. |

2.4 Semantic Frames

We perform the sentence structure analysis based on frame semantics [15], expecting to capture semantically important part in a sentence. Although there are several approaches on the semantic structure and available corpus, we adopt FrameNet frames since FrameNet uses self-explanatory frame and role names. Each sentence is analyzed as one or more frame(s) composed of a main predicate and associated roles.

The FrameNet project [2] defines the frames that organize the semantic information in a sentence. Each frame consists of the target predicate and a set of slots (frame elements) with corresponding semantic roles. FrameNet release 1.2 (June 2005) defines 608 frames with the 7,389 predicates. Figure 2 shows a frame example from the FrameNet.

Frame: Taking_sides

A *Cognizer* has a relatively fixed positive or negative point of view towards an *Issue* (or a *Side* in a debate concerning an *Issue*).

Example Sentence:

In interviews, it seems like everyone is completely *against* this expenditure.



Figure 2. Frame Example from FrameNet

In our work, we perform frame analysis for all verbs in a sentence. A sentence is analyzed in several ways for each verb occurrence.

3. Data Preparation

In order to train an automatic classification and recognition system, we require appropriately annotated material. A manual annotation scheme was developed in collaboration between USC-ISI and the University of Pittsburgh. It was deployed through an iterative, trial and error process using coders working in Pitt's Qualitative Data Analysis Program (QDAP) during the summer and fall of 2005. We asked the annotation of the same document to at least two coders in parallel. We compute the inter-human agreement by Cohen's kappa coefficient [10] and F-measure. Cohen's kappa is an agreement score considering chance agreement, and F-measure is a traditional method to average precision and recall. We compute precision and recall assuming one coder's annotation is a real answer, hence F-measure means the ratio of matches over the other's annotation.

Initially, the QDAP coders employed ATLAS.ti¹, a commercial off-the-shelf qualitative data analysis application. The QDAP coders' existing familiarity with the application of subtopic codes to mercury rulemaking public comment texts using ATLAS.ti enabled the achievement of reasonably high levels of inter-rater reliability (Cohen's kappa coefficient is 0.7 and F-measure is 0.67). However, this annotation scheme was not satisfactory in its ability to reliably capture the linkages that define the argument structure. While ATLAS.ti does allow coders to assign hyperlinks to text segments, it was cumbersome and hence unreliable as a means for building the kinds of complex links necessary for the automatic system.

In order to capture the argument structure of the comments, we developed a java-based annotation tool ("ERuleClient") through which QDAP coders could more easily annotate both the subtopic codes and argument linkage structures within a given comment. One of the challenges in the coding lab involved moving from a familiar proprietary coding system to a custom built system. Since the ERuleClient was ergonomically quite different for coders who had become fluent in coding using ATLAS.ti. A comparison of the coders' annotation from the first iteration with the new ERuleClient to the previous coding in Atlas.ti shows a drop in the reliability of the application of the subtopic codes (Cohen's kappa is 0.51 and F-measure is 0.63).

Table 3. Annotation on long documents

Code	Human1	Human2	Match	Kappa	F-measure
Gov. Resp.	4	8	3	0.50	0.50
Economic	36	64	29	0.54	0.58
Legal	10	14	6	0.49	0.50
Health	49	52	44	0.86	0.87
Science	35	118	20	0.18	0.26
Policy	194	158	125	0.58	0.71
Technology	59	79	39	0.51	0.57
Environment	52	50	22	0.37	0.43
Pollution	191	121	97	0.48	0.62
Total	630	664	385	0.50	0.60
Arg. Struct.	Human1	Human2	Match	F-meas.	F-measure w/ type
Root (Restate)	19 (4)	11 (2)	9	0.60	0.53
Subroot	56	52	20	0.37	0.37
Leaf	145	234	90	0.47	0.47

Table 4. Annotation on short emails

Code	Human1	Human2	Match	Kappa	F-measure
Gov. Resp.	17	25	13	0.57	0.62
Economic	19	24	17	0.76	0.79
Legal	7	14	7	0.65	0.67
Health	49	64	48	0.78	0.85
Science	13	19	10	0.59	0.62
Policy	57	48	42	0.71	0.80
Technology	17	17	16	0.93	0.94
Environment	27	43	23	0.57	0.66
Pollution	85	99	80	0.72	0.87
Total	291	353	256	0.70	0.80
Arg. Struct.	Human1	Human2	Match	F-meas.	F-measure w/ type
Root (Restate)	15 (6)	17 (4)	15	0.94	0.94
Subroot	33	32	24	0.74	0.74
Leaf	31	30	23	0.75	0.75

In subsequent rounds of testing using the ERuleClient tool, we made updates on the fly which significantly improved the coders' satisfaction with the tool. Among the many changes was the addition of a visualization option to allow coders to review the argument trees that they were creating using the tool. Despite the advances in the software, the agreement scores between coders did not increase. It was likely attributable to the length of texts to be annotated. As we selected longer texts (average 107 sentences per text) for review, in hopes of capturing more interesting argument structures, the task itself became more complex. Table 3 shows the agreement on subtopic code and argument structure. For the argument structure agreement, we consider the restating *root* or *subroot* (child linked with the relation of "restate") as same as the *root* (or *subroot*). Among the children of the *roots* showing agreement, *subroot* and *leaf* agreement is computed. "F-

¹ <http://www.atlasti.com>

measure with type” denotes the agreement not only on the role in the structure (“root”, “subroot”, and “leaf”) but also on the link type (“support the regulation”, “oppose the regulation”, and “propose a new idea” for *root*; “support” and “oppose” for *subroot* and *leaf*).

In the most recent round of testing, the PIs selected email texts (average 10 sentences per text). Table 4 shows significantly improved scores for both subtopic code application and argument structure annotation.

Since we require enough data for training and testing, we use the data from each round fully or partially. The specific usage is described in Section 4.

4. System

For each module described in Section 2, the individual system was implemented using the data in Section 3.

4.1 Subtopic Codes

4.1.1 Implementation

To categorize the subtopic into predefined codes, we interpreted a problem as a classification of yes or no, for each subtopic code. Since we assumed that the codes were independent of each other, we performed each separate classifier and assigned all the subtopic codes (from none to multiple codes) given a sentence.

We built a classifier using a support vector machine (SVM) [34]. SVM is a machine learning method widely used in classification problems showing sound performance in many applications. It finds a hyper-plane that separates the positive and negative training examples with a maximum margin in the vector space.

We used annotated texts from every round: 118 documents as a training set and 22 documents as a test set. The training set of 118 documents was annotated by one or more annotators and the total instances of annotations were 274 documents. We used all 274 instances to obtain enough training data expecting that we could assign more confidence on features when they showed agreement between annotators.

The SVM-Light² implementation was adopted using the following semantically oriented features:

-
- Lexeme of word
 - Bigram including stopwords
 - Bigram excluding stopwords
 - Named entity obtained by BBN Identifier³ [5]
 - Named entity label obtained by BBN Identifier
 - Synonyms of the first sense of word in the WordNet 2.0 [14]
-

4.1.2 Evaluation

Table 5 shows the system performance, agreement with the human annotation, for subtopic code detection. Each code showed different performance: for example, “Government Responsibility” was not common for considering its broad

definition, so it ended in low agreement, but “Technology” or “Health” achieved high agreement. The overall agreement is comparable to the human agreement in Table 3 and Table 4.

Table 5. System Performance on Subtopic Code Detection

Code	Human	System	Match	Kappa	F-measure
Gov. Resp.	17	14	4	0.23	0.26
Economic	38	25	18	0.54	0.57
Legal	19	11	7	0.45	0.47
Health	80	106	73	0.73	0.78
Science	46	33	21	0.49	0.53
Policy	133	148	102	0.60	0.73
Technology	26	20	19	0.82	0.83
Environment	40	48	17	0.32	0.30
Pollution	119	116	76	0.52	0.65
Total	518	521	337	0.52	0.65

4.2 Argument Structures and Opinions

For a hierarchical argument structure, first we extracted *root* by a classification method, second, segmented a document and selected a *subroot* from each segment. Next, we defined the linkage and relationship of *support*, *oppose*, and *restate* and assigned the final opinion on the root with one of *support the regulation*, *oppose the regulation*, and *propose a new idea*.

4.2.1 Main Root Identification

Root is the most important part of text containing the writer’s main claim over the regulation. Since multiple *roots* can exist, we defined the problem as a classification into root or not, given all sentences in a document. We used 105 documents annotated by multiple coders (in total, 173 documents of 8,464 sentences) for training. The SVM classifier was applied using the following features:

Word: Words in a sentence excluding stopwords.

Bigram: All pairs of consecutive words in a sentence.

Word’s Stem: Stem words obtained by Porter’s stemmer⁴.

Word frequency in the Summarized Proposal: Based on the assumption that people mention the proposed rule more in the main root, we computed the frequency ratio for each word from the proposal summary. We only considered verb, noun, adjective, and adverb.

Subjectivity: The subjective sentences tend to be a *root* sentence in commentary texts while other objective sentences are supporting reasoning of the *root*. We obtained manually annotated corpus for opinions or emotions, Multi-Perspective Question Answering Corpus⁵ (version 1.1), described in [36]. We extracted all words appearing in the subjective and objective sentences respectively, and applied a Naïve Bayes classifier [25] to compute the subjectivity score for a sentence as follows:

² <http://svmlight.joachims.org/>

³ <http://www.bbn.com>

⁴ <http://www.tartarus.org/~martin/PorterStemmer/>

⁵ <http://www.cs.pitt.edu/~wiebe/pubs/pub1.html>

$$CM = |\log(p(subjective) + \sum_{i=1}^n \log(p(w_i | subjective))) - (\log(p(objective)) + \sum_{i=1}^n \log(p(w_i | objective))))|$$

where $p(subjective)$ is a probability of subjective sentences, $p(w_i | subjective)$ is a probability of the i^{th} word's occurrences of total n words in subjective sentences, and same as for $objective$.

Position: Especially in well-written texts, the $root$ sentence is highly related to the position in text. We indicated a position with three values: paragraph position, sentence position in a paragraph, and relative sentence position in a paragraph.

- *Paragraph position:* The position of a paragraph including a given sentence that is defined as the order of the paragraph in text.
- *Sentence position in a paragraph:* The number representing the order of the sentence in a paragraph.
- *Relative sentence position in a paragraph:* Since the paragraph size is different, the sentence position is represented as a relative position in a paragraph scaled to the interval [0,1].

Cue Phrase: Several cue phrases were utilized.

Please | The EPA should | You should | I hope you | I hope that you | I suggest | Do | In conclusion | In summary | I (we) support | I (we) oppose | I (we) request | I (we) urge | I (we) encourage |

Subtopic code “Policy”: The subtopic code output from Section 4.1 was adopted as a feature. 11% of sentences whose topic was “policy” were *roots* in the training set, and the binary value was used to signal if the sentence covered the subtopic “policy”.

Named Entity: Named Entity (organization, person, and location) recognized by BBN Identifier was used.

4.2.2 Subroot Identification

Subroot is similar to *root*, in terms that it expresses more important and informative material. We simplified the task, assuming that text was a sequence of subtopic segments. We segmented text into several subtopic groups and selected the most important sentence from each segment.

The subtopic segmentation was performed using Hearst's TextTiling [17], which utilized lexical co-occurrence and distribution. To obtain more concentrated and concrete subtopic group, we used a smaller token-sequence size 6 than the default size 20 when computing the similarity between adjacent groups of token-sequences.

To define a single important sentence from a segment obtained, we applied SVM ranker [18] to each segment. We used the same training set as in Section 4.2.1, but extracted the subtree of *subroot* and *leaves* (2,654 sentences). We compared the score within a segment, and selected the one ranked as highest of all sentences in the segment.

The same features in the root identification task in Section 4.2.1 were applied but the position features were altered slightly. Instead of position within a whole text, the relative position within a segment was used.

4.2.3 Link and Link type Identification

After obtaining *root* and *subroot*, we linked all the other topic units (topic-assigned sentences) in the segment to the *subroot*, and linked all the *subroots* to the *root*. When there were more than one *roots* in text, we chose the most semantically similar and locally closest root.

The similarity between topics was obtained by computing cosine similarity. The cosine similarity metric is widely used in information retrieval [29] between documents, but we computed it between two target sentences. The similarity between sentence S_1 and S_2 was defined as follows:

$$\text{Sim}(S_1, S_2) = \frac{\sum_i w_{i,S_1} w_{i,S_2}}{\sqrt{\sum_i w_{i,S_1}^2} \sqrt{\sum_i w_{i,S_2}^2}}$$

$$w_{i,S} = tf_{i,S} \log\left(\frac{N}{sf_i}\right)$$

where $w_{i,S}$ is a weight of term i in sentence S . The term weight is defined as a *tfidf* for each word and bigram. Since this is about the similarity between sentences, *idf* is replaced by “inverse sentence frequency” that counting the frequency in each sentence. $tf_{i,S}$ is frequency of the term i in S , and sf_i is the number of sentences containing the term i , and N is the total number of sentences in text.

To assign a *linktype* (support, oppose, or restate) to each link, we searched “restate” and “oppose” while setting “support” as the default type, since the writers in our domain did not mention contradicting issues a lot.

Restate Link: We define the relation “restate” to signal restating or paraphrasing the same contents. As described above, we computed the cosine similarity between parent and child sentences, and assigned “restate” when the similarity was larger than the empirically found threshold (similarity > 0.15).

Other than between parent and child, the similarity was also computed between pairs of *roots*, and if they were similar, then one *root* (having lower probability for a *root*) was linked to the other *root* as a child using the *linktype* “restate”.

Oppose Link: At this point, only simple cue phrases were used. When two topics were not similar at all and the child contains “even if, even though, although”, we assigned it as “oppose”.

4.2.4 Root type (opinion) Classification

We classify the opinion into three categories (*support the regulation*, *oppose the regulation*, and *propose a new idea*) based on the content of *root* sentences. As most previous research, the positive and negative expressions were checked but we considered semantic units rather than words within a fixed size window or a syntactic clause or phrase.

Each sentence was analyzed into a list of frame elements described in Section 4.3, and “Topic score” and “Polarity score” were computed for each frame element. “Topic score” was defined as a measure of relatedness to the given proposal and “Polarity score” as a measure of polarity of positive and negative expressions as follows:

* *Topic Score*: The sum of each word’s frequency in the proposed rule summary.

* *Polarity Score*: From the opinion annotated corpus [36], positive and negative expressions were extracted, and naïve bayes classifier was built with stem words of polar expressions.

$$CM = \log(p(\text{positive})) + \sum_{i=1}^n \log(p(w_i | \text{positive})) \\ - (\log(p(\text{negative})) + \sum_{i=1}^n \log(p(w_i | \text{negative})))$$

Based on these scores, the final *roottype* (opinion) was determined by simple heuristically derived rules. The detailed procedure is explained in the following:

Given a root sentence,

1) *Identify frame elements described in Section 4.3.2*

2) *For each Frame Element (FE):*

- build a 2-tuple (P, T) where P is polarity score and T is topic score

- Sum tuples into two categories: polarity for topic (topic, polarity score) and polarity for something else (other, polarity score)

3) *For a main predicate verb of the sentence:*

compute polarity score

4) *determine the final roottype by the following rules*

Predicate (negative) => oppose

Predicate (positive) + FE (topic) => support

Predicate (Neutral) + FE (topic, positive) => support

Predicate (Neutral) + FE (topic, negative) => oppose

Predicate (Neutral) + FE (other, positive) => propose

Predicate (Neutral) + FE (other, negative) => oppose

4.2.5 Evaluation

To evaluate the argument structure, first, we compared the *root* with the human annotated *root*, second, checked the *subroots* given the agreed *root*. Since the roots having the same claim were linked with “restate”, if either of *root* or *restating root* matched then we considered it as agreement. Table 6 shows the performance of our system on the argument structure. When we consider the low agreement on the argument structure in long documents (Table 3), the system performance is encouraging although we believe there is space for improvement.

Table 6. System Performance on Argument structure

Type	Human	System	Match	F-measure
Root (restate)	33(7)	22(4)	15	0.55
Subroot	29	45	24	0.65

The agreement in *link type* and *root type* is highly restricted to the *root* and *subroot* agreement since the *link type* is determined for the given parent-child link in the previous step. When the system agreed on the parent-child relations with the human annotation, the system showed the perfect human-system agreement on the *linktype*, where all links had “support” relations (Note that we already considered “restate” relation for the argument element

detection of *root* and *subroot*). Since human coders rarely found “oppose” link in text and they did not agree on the parent-child relations in those cases, it was hard to evaluate the proper agreement on *linktype* between humans and with system, rather we could conclude most links “support” the parent in our domain.

Table 7 shows the accuracy on *roottype* classification. As a baseline to compare, we computed only “polarity score” for a sentence and determined the type (“support the regulation” or “oppose the regulation”). “3-type classification” includes “propose a new idea” which generated more disagreement in human annotation.

Table 7. System Performance on Root Type Classification

Task	Baseline	System
3-type classification	N/A	0.60
2-type classification	0.42	0.77

4.3 Semantic Frames

Focusing on the frames of verbs, we selected *main verbs* from a sentence parsed with the Charniak parser⁶ and chose all verbs having a path of $S-VP+-VB^*$ from a root node S^7 . For example, in the text “The present administration has shown inadequate determination to maintain present standards, or to raise them where justified by cost and benefit analysis.”, we extracted “shown”, “maintain”, “raise”, and “justified” and provided the frame structure for all four predicates.

We selected 120 verb lists from our training data (used in subtopic code classifier in Section 4.1), and searched FrameNet to find 98 corresponding frames including 191 roles. All predicate targets associated with these frames were extracted for training and testing. We obtained 37,764 annotated sentences of the training set, 3,870 sentences of the development set, and 4,745 sentences of the test set.

A shallow semantic parser was implemented based on [21] and [22] using Maximum Entropy models [4]. Given a sentence with a predicate, the frame name was assigned first, and then frame elements were identified and appropriate roles were found for each element.

4.3.1 Frame Classification

The frame classifier was implemented with three feature sets: *lexical unit* (lexeme of word), *lexical type* (verb, noun, or adjective) of the predicate target, and *subcategorization*. *Subcategorization* is defined as a parsing rule that expands the VP (verb phrase) of the predicate verb.

4.3.2 Frame Element Identification

To find the frame element of a sentence, we classified each constituent from a parse tree as being a Frame Element or not. A MaxEnt classifier was implemented using many syntactic and

⁶ <http://www.cs.brown.edu/people/#software>

⁷ These are POS (Part Of Speech) tags defined in Penn Treebank (<http://www.cis.upenn.edu/~treebank/>). S is for sentence, VP is for verb phrase, and VB* is for all verb forms including VB, VBD, VBG, VBN, VBP, and VBZ.

semantic features, and most features were adopted from [22] and [3] (shown in Table 8).

Table 8. Sets for Frame Element Identification

<i>Target</i> : Target word
<i>Lexunit</i> : Lexeme of target word + Target type (verb, noun, adjective)
<i>Path</i> : Path from constituent to target word in syntactic parse tree
<i>S Path</i> : Path from constituent to S of target word
<i>Head</i> : Head word of constituent
<i>Phrase Type</i> : Phrase type of constituent in parse tree (ex. NP, VP)
<i>Logical Function</i> : Governing phrase type (S or VP) of NP
<i>Position</i> : Relative position of constituent to the target word
<i>Voice</i> : Active or Passive voice of target phrase
<i>First Word</i> : First word of constituent
<i>First POS</i> : Part of Speech tag of first word of constituent
<i>Last Word</i> : Last word of constituent
<i>Last POS</i> : Part of Speech tag of last word of constituent
<i>Left Head</i> : Headword of left sibling constituent
<i>Right Head</i> : Headword of right sibling constituent
<i>Named Entity</i> : Named Entity tag of constituent
<i>Head POS</i> : Part of Speech of headword of constituent
<i>Partial Path</i> : Path when constituent is under the same “S” in parse tree
<i>S Count</i> : Number of “S” tags from constituent to target in parse tree
<i>Subcategorization</i> : List of constituent labels under the VP of target

4.3.3 Role Labeling

With identified frame elements, non-overlapping frame element lists were constructed by selecting constituents having higher probability as a frame element when there was overlap between identified frame elements. Role tagging was performed with features including sentence-wide features, and all feature sets are described in Table 9.

Table 9. Feature Sets for Role Tagging

<i>Target</i> : Target word
<i>Lexunit</i> : Lexeme of target word + Target type (verb, noun, adjective)
<i>Head</i> : Head word of constituent
<i>Phrase type</i> : Phrase type of constituent in parse tree (ex. NP, VP)
<i>Logical function</i> : Governing phrase type (S or VP) of NP
<i>Position</i> : Relative position of constituent to the target word
<i>Voice</i> : Active or Passive voice of target phrase
<i>Order</i> : Relative position of frame element in a sentence
<i>Syntactic Pattern</i> : Pattern generated from target word, phrase type, and logical function in a sentence
<i>Previous Class</i> : Role of n th previous constituent

4.3.4 Evaluation

Evaluation on frame analysis was performed on a held-out test set from FrameNet as well as a test set of eRulemaking comments. The results are shown in Table 10 and Table 11. Because of longer and more complicated sentence structures in our domain data, which are different from the structures from FrameNet, the performance dropped by 6% in frame element identification and tagging.

Table 10. Evaluation on test set from FrameNet

Process	Prec.	Recall	F-score
Frame Classification	Accuracy: 0.94		
Frame Element (FE) Identification	0.92	0.73	0.81
Role Tagging given FE	0.84	0.81	0.82
FE identification + Role Tagging	0.75	0.60	0.66

Table 11. Evaluation on test set from eRulemaking data

Process	Prec.	Recall	F-score
Frame Classification	Accuracy: .77		
FE identification + Role Tagging	0.67	0.55	0.60

4.4 Integration

The individual output of each module is integrated to a collective result for all comments. Figure 3 shows the combined result of the test set, and it is an example of the summarized output of multi-texts, which will be provided to rule-writers. All topics were summed up for each opinion category as supporting reasons. When the *linktype* is “oppose”, the topic is added as opposite of the root’s opinion, for example, if the *roottype* is “oppose the regulation” and the child topic is linked with “oppose”, then the topic is summed to “support the regulation”.

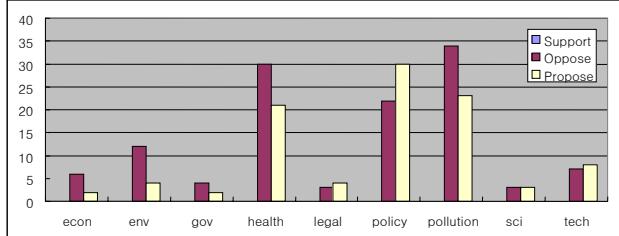


Figure 3. Integrated Output of subtopic code, argument structure, and opinion for test set

Frame: Request - ask, demand, urge
Speaker [I, We, the Center, the Council]
Addressee [EPA, you]
Message
[to withdraw its rulemaking at the present time to implement strong controls on mercury emissions from coal-fired plants return to prior analyses and reduce the SOX cap to 2 million by 2009]
Frame: Removing - eliminate, evict, remove, take, withdraw
Agent [EPA, the Center]
Source [off the 112 (c) list, its rulemaking]
Theme [Mercury Pollution, power plants, the annual testing requirement]
Frame: Reasoning - demonstrate, prove, show
Arguer [EPA, EPA's own analysis, Science, a comprehensive study]
Content
[the present administration that such standards were “appropriate and necessary” that there is questionable basis to regulate mercury emissions from power plants]

Figure 4. Excerpts from Integrated Frame Output for test set

To provide a numerical indication showing agreement in the final graphical output, we computed the cosine similarity between human and system. The term weights were defined as topic frequency per file, subtopic code, and *roottype*. The value between human and system was 0.48, compared to the value 0.67 between two humans.

Figure 4 shows the part of summed output of frame analysis. This shows semantic frames including semantic roles and real instances found in the comments.

5. Conclusion

We have described our system to extract various aspects of information from texts including the annotation process. We are planning to investigate a way to improve the individual step by defining the codes and the manual annotation task more clearly and by using more generalized pattern-based features. For a prototype system to be provided to rule-writers, we will conduct more analyses of the trends and new aspects of future public comment. Further, we plan to combine this work with that of our colleagues at CMU on near-duplicate detection and to create a system that performs multi-aspect analysis of rulemaking comments and provides a useful review tool for rule-writers.

6. ACKNOWLEDGMENTS

The researchers wish to acknowledge the EPA for providing the datasets on which this report is based. This work was supported by NSF grants IIS-0429293 and IIS-0429360.

Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the National Science Foundation.

7. REFERENCES

- [1] Altman, D., *Practical Statistics for Medical Research*. Chapman and Hall, 1991.
- [2] Baker, C.F., Fillmore, C.J., and Lowe, J.B., The Berkeley FrameNet project. In *Proceedings of COLING-ACL*, Montral, Canada, 1998.
- [3] Bejan, C.A., Moschitti, A., Morarescu, P., Nicolae, G., and Harabagiu S., Semantic Parsing based on FrameNet. In *Proceedings of ACL-SENEVAL workshop on the Evaluation of Systems for the Semantic Analysis of Text*, Barcelona, Spain, 2004.
- [4] Berger, A., Pietra D., and Pietra, V.D., A Maximum Entropy Approach to Natural Language. *Computational Linguistics*, 22(1), 1996.
- [5] Bikol, D., Schwartz R., and Weischedel, R. M., An Algorithm that Learns What's in a Name. *Machine Learning*, 34 (1-3), pp. 211--231. 1999.
- [6] Burstein, J., Marcu, D., Andreyev, S., and Chodorow, M., Towards automatic classification of discourse elements in essays. In *Proceedings of the 39th annual Meeting on Association for Computational Linguistics*, Toulouse, France, 2001.
- [7] Brill, E., Some Advances in Transformation-Based Part of Speech Tagging, In *Proceeding of the 12th National Conference on Artificial Intelligence*, Seattle, WA. 1994.
- [8] Coglianese, C. The Internet and Citizen Participation in Rulemaking. *I/S* 1(1): 33-57. 2005.
- [9] Coglianese, C. E-Rulemaking: Information Technology and the Regulatory Process. *Administrative Law Review* 56(2): 353-402. 2004.
- [10] Cohen, J., A coefficient of agreement for nominal scales. *Education and Psychological Measurement*, 43(6):37—46. 1960.
- [11] Crags, R. and Wood, M., Evaluating Discourse and Dialogue Coding Schemes. *Computational Linguistics*, 31(3),pp. 289-295, 2005.
- [12] Cuéllar, M.F., Rethinking Regulatory Democracy. *Administrative Law Review* 57(2): 411-499. 2005.
- [13] Emery, F. Emery, A., A Modest Proposal: Improve E-Rulemaking by Improving Comments. *Administrative and Regulatory Law News*, 31(1): 8-9. 2005.
- [14] Fellbaum, C., *An Electronic Lexical Database*, The MIT press. 1998.
- [15] Fillmore, C.J., Frame Semantics and the Nature of Language. *Annals of the New York Academy of Science Conference on the Origin and Development of Language and Speech*, 280: 20-32. 1976.
- [16] Fleiss, J., *Statistical Methods for Rates and Proportions*. John Wiley & Sons, 1981.
- [17] Hearst, M., TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages. *Computational Linguistics*, 23 (1), pp. 33--64. 1997.
- [18] Joachims, T., Optimizing Search Engines Using Clickthrough Data, In *Proceeding of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*, ACM, Edmonton, Alberta, Canada, 2002.
- [19] Kim, S. and Hovy, E., Determining the Sentiment of Opinions. In *Proceedings of COLING*, Geneva, Switzerland, 2004.
- [20] Krippendorff, K., *Content Analysis: An Introduction to Its Methodology*. 2nd ed. Sage, Beverly Hills, CA. 2004.
- [21] Kwon, N., Fleischman, M.B, and Hovy, E., FrameNet-based Semantic Parsing using Maximum Entropy Models. In *Proceedings of COLING-04*, Geneva, Switzerland. 2004.
- [22] Kwon, N., Fleischman, M., and Hovy, E., SENSEVAL Automatic Labeling of Semantic Roles Using Maximum Entropy Models. In *Proceedings of ACL-SENEVAL workshop on the Evaluation of Systems for the Semantic Analysis of Text*, Barcelona, Spain, 2004.
- [23] Lau, G.T., Law, K.H., and Wiederhold, G., A Relatedness Analysis Tool for Comparing Drafted Regulations and Associated Public Comments. *I/S* 1(1): 95-110. 2005.
- [24] Mann, W.C. and Thompson, S., Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8(2):243-281. 1988.
- [25] Mitchell, T., *Machine Learning*, McGraw-Hill. 1997.

- [26] Office of Management and Budget, Bush Administration Cuts Regulatory Cost Growth by 70%, Press Release available at: <http://snipurl.com/kofg>. 2005.
- [27] Possin, K., *Critical Thinking*. The Critical Thinking Lab. 2002.
- [28] Pang, B., Lee L., and Vaithyanathan, S., Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of EMNLP*, Philadelphia, PA. 2002.
- [29] Salton, G., Wong, A., Yang, C.S., A Vector Space Model for information Retrieval. *Communications of the ACM*. 18(11):613-620. 1975.
- [30] Shulman, S.W., E-Rulemaking: Issues in Current Research and Practice. *International Journal of Public Administration* 28: 621-641. 2005.
- [31] Shulman, S.W., The Internet Still Might (But Probably Won't) Change Everything. *I/S* 1(1): 111-145. 2005.
- [32] Teufel, S. and Moens, M., Discourse-level Argumentation in Scientific Articles: Human and Automatic Annotation. In *Proceedings of ACL workshop on Towards Standards and Tools for Discourse Tagging*, College Park, MD. 1999.
- [33] Turney, P., and Littman, M., Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transactions of Information Systems (TOIS)* 21(4):315-346. 2003.
- [34] Vapnik, V. N., *The nature of Statistical Learning Theory*, Springer. 1995.
- [35] Wiebe, J., Wilson, T., Bruce, R., Bell, M., and Martin, M., Learning Subjective Language. *Computational Linguistics* 30(3):277-308. 2004.
- [36] Wilson, T., Wiebe, J., Annotating Opinions in World Press. In *Proceedings of SIGdial-03*. Sapporo, Japan, 2003.
- [37] Wilson, T., Wiebe, J., and Hoffmann, P., Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Proceedings of HLT-EMNLP*, Vancouver, Canada. 2005.
- [38] Yang, H. and Callan, J., Near Duplicate Detection for eRulemaking. In *Proceedings of the Sixth National Conference on Digital Government Research*, Atlanta, GA. 2005.
- [39] Yu, H. and Hatzivassiloglou, V., Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. In *Proceedings of EMNLP-03*, Sapporo, Japan, 2003.

Automatically Labeling Hierarchical Clusters

Pucktada Treeratpituk

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
puck@cs.cmu.edu

Jamie Callan

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA

callan@cs.cmu.edu

ABSTRACT

Government agencies must often quickly organize and analyze large amounts of textual information, for example comments received as part of notice and comment rulemaking. Hierarchical organization is popular because it represents information at different levels of detail and is convenient for interactive browsing. Good hierarchical clustering algorithms are available, but there are few good solutions for automatically labeling the nodes in a cluster hierarchy.

This paper presents a simple algorithm that automatically assigns labels to hierarchical clusters. The algorithm evaluates candidate labels using information from the cluster, the parent cluster, and corpus statistics. A trainable threshold enables the algorithm to assign just a few high-quality labels to each cluster. Experiments with Open Directory Project (ODP) hierarchies indicate that the algorithm creates cluster labels that are similar to labels created by ODP editors.

Categories and Subject Descriptors

H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing—*Linguistic processing*.

General Terms

Algorithms, Performance, Experimentation.

Keywords

Document hierarchy, cluster labeling.

1. INTRODUCTION

Government agencies, like most modern organizations, must often quickly organize and analyze large amounts of textual information. Our research is motivated by notice and comment rulemaking, in which U.S. regulatory agencies are required to consider comments submitted by the general public [11]; when an agency receives hundreds of thousands of unique comments and edited form letters during a short period of time, quickly organizing them and identifying the issues raised is a significant challenge. However, the problem of organizing large amounts

of text for rapid analysis subsumes notice and comment rulemaking. The U.S. National Archives and Records Administration (NARA) and other government agencies face similar problems. Search engines such as FirstGov¹ have the problem on a smaller scale – given a set of texts, how to quickly organize them and describe their contents.

Hierarchical organization is popular because it represents information at different levels of detail and is convenient for interactive browsing (e.g., Yahoo! [17], the Open Directory Project [13]). At the top of the hierarchy, the collection is organized into a few general categories; as a person descends the hierarchy, she gets greater detail about increasingly specific categories. Typically each document cluster is assigned a descriptor that describes the documents it contains. One goal of hierarchical clustering is to improve the users' ability to browse the collection, so it is very important that the hierarchy has good cluster descriptors. These descriptors can be either category labels, as in Yahoo! Directories, or lists of topical terms.

There is considerable prior research on hierarchical clustering algorithms and their applications in information retrieval and data mining research. However, less attention has been paid to creating good cluster descriptors. Cluster descriptors created automatically often either fail to provide a comprehensive description of the cluster, or consist of lists of terms from which a person must infer a general description.

This paper proposes a simple algorithm that automatically assigns concise labels to hierarchical clusters. The algorithm combines statistical features of the cluster, the parent cluster, and the corpus into a descriptive score. The algorithm is based on the hypothesis that by comparing the word distribution from different parts of the hierarchy, it should be possible to assign appropriate labels to each cluster in the hierarchy.

The rest of this paper is organized as follows. Section 2, describes the hierarchical clustering and hierarchical clustering labeling task. It also discusses the characteristic of a good label for hierarchical clusters. Section 3 presents previous research on cluster labeling and related tasks. Section 4, describes the proposed labeling algorithm. Section 5 presents experimental results. Section 6 summarizes our finding and offers suggestions about possible future improvements.

¹ <http://www.firstgov.gov/>

2. LABELING HIERARCHICAL CLUSTERS

Hierarchical clustering partitions a document collection into a small number of clusters, and each cluster is further partitioned into subclusters in a recursive manner. Hierarchical clusters can be constructed by *agglomerative* methods that start with each document in its own cluster and then repeatedly group similar clusters into broader clusters; or by *divisive* methods that start with all documents in one cluster and then repeatedly divide each cluster into more detailed subclusters.

In labeling hierarchical clusters, one assumes the existence of a hierarchy of document clusters. The task is to assign a good descriptor to each cluster node in the hierarchy. The most common cluster descriptors are either concise labels, or lists of terms and phrases. For example, a cluster of documents about natural language processing might be described by the label “natural language processing” or the list of terms “tag, text, linguist, lexicon, corpus, tagger, word, syntax, grammar.” A list of terms is often less useful than a single category label, because it requires the user to infer the concept implied by the terms. However, a list of terms is the most common choice for labeling clusters automatically because it fails gracefully; a person can often infer the general description even when a few of the selected terms are poor choices.

Our goal is an algorithm that selects concise cluster labels that are similar to what a person might create manually. A good descriptor for a cluster should not only indicate the main concept of the cluster, but also differentiate the cluster from its siblings and its parent cluster. Consider a cluster of “neural network” documents under a broader “AI” cluster. The parent cluster “AI” may also have “machine learning” and “fuzzy logic” child clusters. In another context, a label “computer science” might be an acceptable descriptor for the “neural network” cluster. However, in the context of this hierarchy, “computer science” does not distinguish the “neural network” cluster from its siblings. The descriptors appropriate for a given set of documents will differ under different hierarchies.

We define the labeling task as follows: Given a cluster of documents in a cluster hierarchy, the goal is to produce appropriate category labels for the cluster. The algorithm is allowed to return a list of plausible labels, ranked by its confidence about how descriptive each label is; the list should be as short as possible. A ranked list of labels is a compromise between a single category label and a list of topical terms.

3. RELATED WORK

Although there has been much research in hierarchical clustering of documents, little focused on labeling the resulting clusters of documents [2][3][4][5][10]. Clustering algorithms can naively label their clusters with the most frequent words in the clusters [2]. Those resulting labels tend to be general, and usually are not good discriminative descriptors. The algorithm may choose to represent its clusters of documents with the documents near each cluster’s centroid. One example of algorithms that use this representation is Scatter/Gather [3], which represents a cluster with a list of documents near the cluster’s centroid and a list of topical terms. The topical terms are the terms with the highest weights in the cluster centroid. In one report the algorithm showed approximately the top 10 topical terms to the users. In addition to the shortcomings, mentioned above, of using lists of

terms as descriptors, this approach does not take into account the hierarchical structure of the cluster hierarchy. The resulting descriptor might be descriptive, but might not discriminate the cluster from its parent or sibling clusters.

There have been attempts to identify cluster labels from word distribution in the hierarchy. Popescul et al. [10], and Glover et al. [4] proposed statistical methods in selecting cluster descriptors, based on the context of the surrounding clusters (parent cluster and sibling clusters). Popescul proposed to use the statistical test χ^2 to detect difference in word distribution across the hierarchy. At each cluster node in the hierarchy, starting from the root, the χ^2 test is used to detect a set of words that is equally likely to occur in any of the subclusters of the current node. Those words are considered to be non-descriptive terms for every subcluster of the current node, and thus are removed from every subcluster. After the χ^2 test is used to remove non-descriptive words from every cluster node, the algorithm labels each cluster with the list of the remaining words at that cluster node ranked by the word frequency.

Glover et al. [4] showed how a simple model based only on terms’ document frequency statistic can be used to select parent, child and self descriptors for document clusters, especially for web pages. The label candidates were extracted from the web page’s content, anchor texts and extended anchor texts. Anchor texts of a web page are the hyperlinked words that link to the web page. Extended anchor texts refer to the words that occur before and after the anchor texts including the anchor texts themselves. Labels were selected and ranked using document frequency and some preset cutoff values. In their experiment, they found that labels extracted from anchor texts and extended anchor texts provide better description than ones extracted from the page’s content. This is because web pages often do not contain words that describe their categories. However, obtaining anchor texts and extended anchor texts is an expensive operation, and requires one to know the hyperlink structure of the World Wide Web. Furthermore, the hyperlink structure is generally not available in non-webpage document collection.

Another research area closely related to cluster labeling is automatic ontology construction. It should be noted that while a ontology hierarchy has well-defined parent-child relationship, such as hypernyms-hyponym and meronymy (part-whole), a document hierarchy of the same collection does not necessarily have to reflect the same parent-child relationship. The higher flexibility in hierarchical structure of a document hierarchy might better serve for task such as browsing.

There have been some works on creating ontology hierarchies using clustering based techniques. These approaches require that the resulting hierarchy be automatically labeled. Caraballo [1] constructed a noun hierarchy of hypernyms automatically from text. The noun hierarchy is constructed using bottom-up clustering approach, grouping nouns based on conjunction and apposition. In order to label each internal cluster, a set of possible hypernyms of every noun in the cluster is extracted from the text using a linguistic pattern. The noun that has the largest number of hyponym relations with the noun in the clusters is assigned as the cluster label.

Pantel et al. [9] automatically assigned label to semantic classes, generated from their clustering algorithm. For each semantic class, a subset of concepts in the class that is most likely to represent the semantic class is selected as class representatives.

These representative concepts are then used to extract label candidates using some lexical patterns. The label candidate with the highest mutual information with the class representatives is assigned as class label.

4. ALGORITHMS

Glover et al. [4] showed that a simple term frequency analysis could predict the labels of document clusters. Their algorithm is based on the hypothesis that a word that is very common in the cluster, but relatively rare in the collection, is likely to be a good cluster descriptor. They selected cluster descriptor candidates based on the following criteria:

$$\text{Candidates} = \{\text{phrase } p \mid DF_C / |C| < \text{maxColPos} \text{ and } DF_S / |S| > \text{minSelfPos}\}$$

where DF_C is the number of documents in the collection that contains the phrase p ("document frequency"), and DF_S is the number of documents in the cluster (the "self cluster" S) that contain the phrase p . $|C|$ and $|S|$ denote the number of documents in the collection and in the self cluster. maxColPos and minSelfPos are thresholds. Phrases that appear more than minSelfPos times per document, on average, in "self cluster" documents, and less than maxColPos times per document, on average, in the collection are considered to be in the label candidate set. Phrases in the candidate set are ranked according to their DF_S values. Every phrase in the collection is considered, without stemming or stopwords removal [4].

There are several limitations to Glover's threshold-based method. First, the performance of the algorithm is sensitive to preset threshold values (maxColPos , minSelfPos), and the optimal thresholds vary between clusters. Second, multiword phrases normally have lower $DF_S / |S|$ values than single words, thus are rarely selected as descriptors. Third, documents often do not contain words that describe their categories, so basing the decision mainly on $DF_S / |S|$ generally does not work well.

Due to these limitations, we propose a more general labeling algorithm that allows us to incorporate more features in selecting the cluster descriptors.

Labeling Algorithm:

First, we assume that the algorithm has access to a general collection of documents E , representing the word distribution in general English. This English corpus is used primarily in selecting label candidates, as explained below.

Given a cluster S and its parent cluster P , which includes all of the documents in S and in the sibling clusters of S , the algorithm selects labels for the cluster S with the following steps:

- 1) **Collect phrase statistics:** For every unigram, bi-gram, and tri-gram phrase p occurring in the cluster S , calculate the document frequency and term frequency statistics for the cluster, the parent cluster and the general English corpus.
- 2) **Select label candidates:** Select the label candidates from unigram, bi-gram, and tri-gram phrases based on document frequency in the cluster and in general English language.
- 3) **Calculate the descriptive score:** Calculate the descriptive score ($D\text{Score}$) for each label candidate, then sort the label candidates by these scores.

- 4) **Calculate the cutoff point:** Decide how many label candidate to display based on the descriptive scores.

Each step is described in more detail, below.

4.1 Collecting Phrase Statistics

For each phrase appearing in the cluster, collect the following statistics: document frequency (DF), and term frequency (TF) with respect to the cluster S , the parent cluster P and the general English corpus E . Document frequency of a phrase p with respect to a cluster C , denoted by DF_C , is the number of documents in the cluster that contain p . Term frequency of a phrase p in a cluster C , denoted by TF_C , is total number of occurrences of p in the cluster.

4.2 Select Label Candidates

Instead of considering every phrase occurring in the cluster, we hypothesize that, although a good descriptor need not occur in the majority of the documents in the cluster, it should occur in at least 20% of the documents in the cluster. Since phrases in general occur less frequent than single words, the selection criteria are slightly different in the case of bigrams and trigrams: The algorithm only considers bigram and trigram phrases that occur in at least 5% of the documents in the cluster. These cutoffs improve the efficiency of the algorithm. Low frequency phrases usually have low weights, thus these thresholds generally don't prune phrases that would be ranked highly otherwise.

Common words (stopwords) are also removed from consideration. The algorithm considers any words that occur in more than 20% of the general English corpus to be stopwords. In the case of the non-unigram phrases, the algorithm considers any phrases that contain only words that occur in more than 30% of the documents in the general English corpus E to be stopwords. This cutoff was chosen conservatively by analyzing the word distribution in general English corpora, trying not to exclude descriptive words.

4.3 Descriptive Score ($D\text{Score}$)

The descriptiveness of a label with respect to a cluster is measured by the descriptive score ($D\text{Score}$). For a phrase p , the descriptive score is based on the features described below.

Normalized Document Frequency ($DF_C / |C|$)

Normalized document frequency is the fraction of the cluster that contains the phrase p .

$$\text{normalized } DF_C = \frac{DF_C}{|C|}$$

In general a label candidate that occurs in more documents in the cluster is expected to be a better descriptor than one that rarely occurs. The algorithm computes *normalized DF* for both the self cluster S and the parent cluster P ($DF_S / |S|$, $DF_P / |P|$). A good descriptor should occur relatively frequent in the parent cluster, but occur very frequent in the self cluster.

TFIDF

This is similar to a traditional *TFIDF* value used in information retrieval.

$$TFIDF_C = TF_C * \log\left(\frac{|C|}{DF_C}\right)$$

As in traditional IR, a phrase with high *TFIDF* value is expected to be important to the cluster, thus possibly is also a good cluster descriptor. The *TFIDF* score favors phrases that appear multiple times per document. The algorithm computes *TFIDF* for both the self cluster and the parent cluster ($TFIDF_S$, $TFIDF_P$).

Rank of *TFIDF*, and *nDF* ($r(TFIDF)$, $r(normalized DF)$)

Four *rankings* are computed for every label candidate based on the features $DF_S / |S|$, $DF_P / |P|$, $TFIDF_S$, and $TFIDF_P$. For example, for $DF_S / |S|$, the algorithm sorts every label candidate according to its $DF_S / |S|$ score; the label with the highest value is assigned rank 1, denoted by $r(DF_S / |S|) = 1$. The label with the second highest score is assigned rank 2, and so on. Tied scores produce tied rankings. A good descriptor is expected to have a relatively high rank of $DF_P / |P|$ and even higher rank of $DF_S / |S|$.

Rank features, e.g. $r(DF_S / |S|)$, convey similar information as their quantitative counterparts, e.g. DF_S . However, rank features may be less sensitive than normalized *DF* and *TFIDF* values, and thus may be more comparable across categories.

Boost in Ranking

Since we hypothesize that a good descriptor probably has a relatively high rank of normalized DF_P (relatively frequent in the parent cluster), and even higher rank of normalized DF_S (very frequent in the self cluster), we measure this boost in ranking of *nDF* with the following measure:

$$\log\left[r\left(\frac{DF_P}{|P|}\right)\right] - \log\left[r\left(\frac{DF_S}{|S|}\right)\right]$$

The algorithm computes the boost in ranking in log-scale because the change in ranking is more significant at the top of the ranking (those which have high document frequency). For example, a label that moves from being the 200th most frequent phrase in the parent cluster to the 100th most frequent phrase in the self cluster, is probably less descriptive than another label that moves from the 100th most frequent phrase in the parent cluster to the 5th most frequent phrase in the self cluster.

In addition to the boost in ranking of normalized document frequency, the algorithm also computes the boost in ranking of *TFIDF*, with the following formula:

$$\log[r(TFIDF_P)] - \log[r(TFIDF_S)]$$

If *TFIDF* is related to the topicality of the phrase, then a good descriptor is expected to have a higher *TFIDF* rank in the self cluster, compared to its *TFIDF* rank in the parent cluster. The algorithm also computes the boost in ranking of *TFIDF* in log-scale for the same reason as in the case of normalized *DF*.

Phrase Length (LEN)

The phrase length is the number of terms in the phrase. While the document frequency feature prefers a shorter phrase to a longer phrase, LEN prefers the longer phrases to shorter ones.

The algorithm combines every feature into one descriptive score with a linear model. Thus the algorithm computes how descriptive a phrase p is as a label for the cluster S, with parent cluster P with the following formula:

$$\begin{aligned} DScore_p &= c_0 + c_1 * LEN \\ &+ c_2 * DF_S / |S| + c_3 * DF_P / |P| \\ &+ c_4 * TFIDF_S + c_5 * TFIDF_P \\ &+ c_6 * r(DF_S / |S|) + c_7 * r(DF_P / |P|) \\ &+ c_8 * r(TFIDF_S) + c_9 * r(TFIDF_P) \\ &+ c_{10} * [\log(r(DF_P / |P|)) - \log(r(DF_S / |S|))] \\ &+ c_{11} * [\log(r(TFIDF_P)) - \log(r(TFIDF_S))] \end{aligned}$$

Each label candidate is sorted by its descriptive score.

The weights of each feature are estimated using linear regression and training data. Linear regression attempts to estimate the expected value of a variable Y given the values of a set of features X_i , by assuming a linear relationship between Y and X_i . Thus, Y can be expressed as linear combination of features Xi:

$$Y = b_0 + \sum b_i X_i + e$$

where e is a random variable residue (error term), with mean zero. The coefficients b_i for all i are optimized so that the sum of the residue square in the training data is minimized.

In order to train the linear regression model, since the correct descriptive score is not known for each label candidate, we have to estimate the descriptive score of a label candidate. We estimate each label candidate's descriptive score based on how much the label overlaps with the correct category label in a set of training data. We define the DScore estimate of a label L , with respect to the correct label CL as:

$$DScore_L^* = \max_{SL \in Synonym(L)} \left\{ \frac{\text{overlap}(SL, CL)}{\max\{\text{len}(SL), \text{len}(CL)\}} \right\}$$

where $\text{overlap}(SL, CL)$ is the number of terms that are shared between SL and CL , and $\text{len}(X)$ denotes the length of X . If the label candidate or a synonym of the label candidate is the same as the correct category label, then the DSscore estimate is 1. The estimation of DSscore that we use to train the linear regression model is only a heuristic value, because many good descriptors would have the DSscore estimates of zero, since they do not overlap with the correct label.

4.4 Cutoff Model

By default, the algorithm displays the 5 labels with the highest descriptive scores as the cluster descriptor. However, we observe that even in a short list of five labels there is generally a big drop-off in the descriptive scores at some point. The big drop-off in the descriptive scores often separates good labels from bad labels. The algorithm can use this information to decide how many labels to display. If the top-ranked label has a very high descriptive score compared to the rest of the label candidates then the algorithm can be very confident that the top-ranked label is the correct descriptor, and thus display only the top-ranked label. On the other hand, if all label candidates have similarly low DScores (only small gaps between each consecutive label), there is less certainty about which labels are best, so more labels are displayed.

The following linear model is used to decide how many labels to display.

$$\begin{aligned}\# \text{Displayed} = & c_0 - c_1 * (DScore_{L1} - DScore_{L2}) \\& - c_2 * (DScore_{L2} - DScore_{L3}) \\& - c_3 * (DScore_{L3} - DScore_{L4}) \\& - c_4 * (DScore_{L4} - DScore_{L5})\end{aligned}$$

The weights in the model are optimized using linear regression and training data. We expect the optimized C_0 to be around 5, while expecting the weights C_1, C_2, C_3, C_4 to be in increasing order.

The same training data used to train the descriptive score model in Section 4.3 can be used to generate training instances for the cutoff model. The cutoff model is trained based on the top-ranked labels produced by the trained descriptive model. For each category in the training data, the 5 labels with the highest predicted descriptive score are determined. The optimal number of labels to display is defined as the rank of the label (from 1 to 5) that has the maximum overlap with the correct category label. If there is a tie, then the label furthest down the list is picked. Thus the cutoff model is trained based on the top-5 predicted descriptive scores from each training category.

5. EXPERIMENTAL RESULTS

A set of experiments was conducted to evaluate the effectiveness of the algorithm at selecting labels. We describe the data and evaluation measures first, followed by descriptions of the experiments and their results.

5.1 Data Collections

The Open Directory Project² was used as a source of documents (web pages), hierarchical organization, and “ground truth” labels assigned by human editors. We randomly sampled 20,462 web pages from the ODP hierarchy to use as background model representing general English (collection statistics). We separately sampled another subset of ODP hierarchy to use as our training and testing ground truth data.

We selected total of 165 subcategories from ODP under 9 categories.

- Computers / artificial intelligence.
- Computers / security.
- Health / alternative.
- Health / medicine.
- Health / medicine / imaging.
- Health / medicine / surgery.
- Health / conditions and diseases.
- Health / conditions and diseases / digestive disorders.
- Business / management.

Every subcategory from the 9 parent categories was included, with the exception of alias subcategories (because those subcategories mainly belong somewhere else in the hierarchy), even ones with common words as labels such as Surgery / General. In total, the constructed hierarchy contains 25,143 web pages. The selected subcategories vary both in depth (between level two and level three with respect to the Open Directory

² <http://www.dmoz.org/>

root) and in number of web pages in each cluster. Figure 1 shows a partial snapshot the document hierarchy.

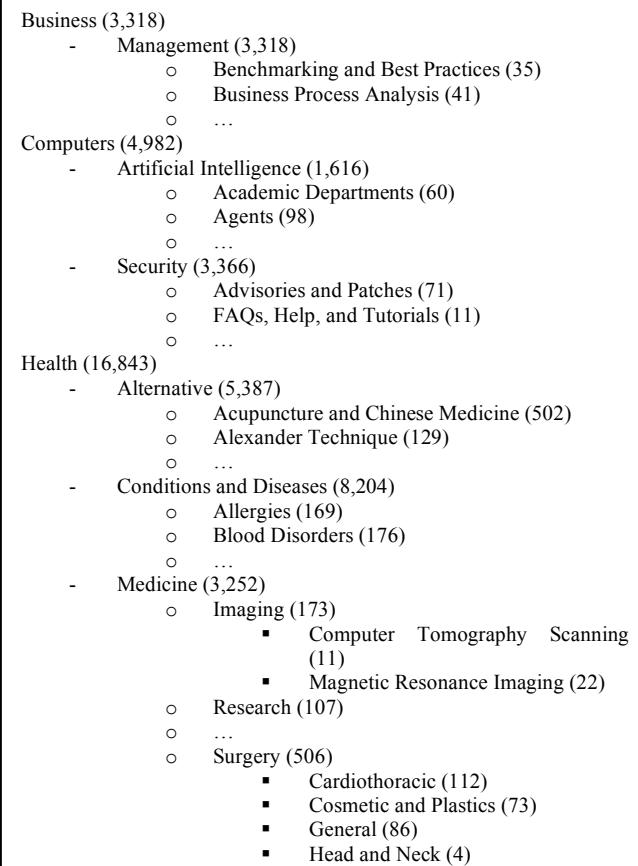


Figure 1. shows subset of the ODP hierarchy used as ground truth data, number of documents in each cluster is showed in the parentheses

5.2 Evaluation Measure

We define the cluster-labeling task as a descriptor-ranking problem. In the evaluation we need to specify our criteria in assessing the quality of the ranking produced. In comparing our labels to the correct ODP label, we use the following two definitions of a correct label: Exact match and partial match.

5.2.1 Definitions of a correct label

For a given category with self-label S , and parent-label P :

Exact Match: A label L is an exact match of the correct label S if there exists a synonym SL of L such that SL is equal to “ S ”, “ $S P$ ” or “ $P S$.” For example, for the category “medical / research,” labels such as “research,” “medical research,” and “research medical” would be classified as exact match labels.

Partial Match: A label L is a partial match of the correct label S if there exists a synonym SL of L such that SL shares a term with “ S ”, “ $S P$ ” or “ $P S$.” For example, for the category “management / business process analysis,” labels such as “business,” “process,” “business management,” “management analysis” would be classified as partial matches.

The synonym list for each word was obtained automatically from WordNet [16]. For both of these definitions of a correct label, we compute the following evaluation measures.

5.2.2 Match at top N results (*Match@N*)

Match@N indicates whether the top N results contain any correct labels. It is a binary indicator, and monotonically increases as N increases.

5.2.3 Precision at top N results (*P@N*)

Precision is computed as the number of labels in the top N results that match the correct categories label divided by N. *P@N* measures the percentage of correct answers that are displayed in ranks 1-N. In general, low precision is undesirable.

5.2.4 Mean Reciprocal Rank (MRR)

Mean reciprocal rank is the mean of the reciprocal of the rank of the first correct label. If the first correct label is ranked as the 3rd label, then the reciprocal rank (RR) is 1/3. If none of the first N responses contains a correct label, RR is 0. RR is 1 if the highest ranked label matches the correct label.

5.2.5 Mean Total Reciprocal Rank (MTRR)

Sometimes there is more than one aspect to a category; for example, the category “acupuncture and Chinese medicine” has two correct aspects, “acupuncture” and “Chinese medicine.” MTRR is similar to MRR, however, instead of considering only the rank of the first correct label as in MRR, MTRR takes into account all correct labels. Of the algorithm ranks “acupuncture” and “Chinese medicine” as the 2nd and the 4th labels, then the TRR (total reciprocal rank) is $\frac{1}{2} + \frac{1}{4} = \frac{3}{4}$ while RR = $\frac{1}{2}$.

Our evaluation methodology is extremely strict because it measures agreement with the single ODP category label selected by the human editor, whereas in fact there might be several equally good category labels. For example, in the category “cardiovascular disorder,” our algorithm might select “heart” and “heart disease” as labels for the cluster, which would be acceptable labels to most human assessors. Our automatic evaluation would judge “heart” and “heart disease” as unacceptable answers. We alleviate some of this problem by accepting synonymous labels, as defined by WordNet synonym lists [16] in our evaluation. However, there are still cases such as

ODP category labels [parent-label/self-label]	#Docs	Labels
artificial intelligence/agents	84	agent, software agent
artificial intelligence/conferences and events	72	conference, artificial intelligence, international conference
artificial intelligence/genetic programming	65	genetic, genetic algorithm
artificial intelligence/philosophy	46	philosophy, mind, science
artificial intelligence/vision	61	vision, compute vision
security/conferences	14	secure conference, conference attend
security/honeypots and honeynets	62	honeypot, attack
security/news and media	73	attack, vulnerable, hack
alternative/apitherapy	18	bee, honey
alternative/ear candling	10	ear candle, ear
alternative/herbs	258	herb, plant
alternative/iridology	18	iridology, iris
alternative/non-toxic living	53	toxic, environmental, safe
alternative/reflexology	95	reflexology, reflexologist
alternative/urine therapy	6	urine
conditions and diseases/chronic illness	53	ill, chronic ill, chronic
digestive disorders/esophagus	35	reflux, heartburn
digestive disorders/pancreas	10	pancreas, pancreatiti
conditions and diseases/food and water borne	75	food, diarrhea
conditions and diseases/musculoskeletal disorders	473	pain, arthritis, joint
conditions and diseases/skin disorders	383	skin, treat
medicine/education	437	continue medical educate, medical educate, medical school
imaging/computer tomography scanning	11	tomography, compute tomography, compute
imaging/x-ray	15	breast cancer, cancer
medicine/reference	119	database, medical subject head, library
surgery/cryosurgery	7	cryosurgery, treat, cryotherapy
surgery/orthopedics	19	orthopaedic, hip
surgery/transplant	36	transplant, transplantation
management/business process analysis	31	business process, business process model
management/management science	430	university, research, paper
management/value based management	8	shareholder value, consult firm, firm

Figure 2. clusters' labels predicted by the descriptive score with cut-off model for 31 categories.

“cardiovascular” and “heart”, which are not actually synonyms, but which most people would consider acceptable substitution labels in the context of “heart disease.”

5.3 Experimental Setup & Results

We evaluated our model on the ground truth ODP data of 165 categories, with a total of 25,143 web pages. Each web page was parsed and all HTML tags, images, and JavaScript were removed in the preprocessing step. Each term was stemmed using Krovetz’s stemmer [6]. No stopwords list was used, because we expected the algorithm to be able to distinguish the collection-specific stopwords from the content words. The goals of our experiments were three-fold. First, we wanted to evaluate the quality of the cluster labels produced by the algorithm, in comparison with the previous technique. Second, we wanted to investigate the performance of the model that uses only rank features. Third, we wanted to investigate how the model performs if the hierarchical cluster is noisy, as would be the case when using a hierarchical clustering algorithm to organize documents.

5.3.1 Performance Comparison

Glover’s threshold-based algorithm was used as the baseline system. The experiment used five-fold cross-validation; in each fold, the training data was used to estimate the optimal parameters for each algorithm. In the case of the threshold-based model, the training data was used to find the optimal threshold values. In our algorithm, the training data was used to learn the weights in the linear model of the descriptive score. This training data was also used to train the cutoff model to predict how many labels to show.

In the training phase we first generated training instance-value pairs for the descriptive score and the feature set training. For each category in the training data, we estimated the DScore for each of its label candidates as described in Section 4.3. We also trained the cutoff model as described in Section 4.4.

The experiment was run on the baseline system and two versions of our algorithm: One with just the descriptive score, and another with both the descriptive score and the cutoff model. Tables 1, 2 and 3 show results for the three algorithms.

Table 1. Match@N with exact, and partial match criteria.

Match@N (exact)	N = 1	N = 2	N = 3	N = 4	N=5
Glover’s	0.27	0.35	0.42	0.46	0.50
DScore	0.36	0.50	0.58	0.62	0.64
DScore + Cutoff	0.37	0.49	0.55	0.55	0.55
Match@N (partial)					
Glover’s	0.39	0.52	0.60	0.64	0.68
DScore	0.53	0.63	0.69	0.72	0.76
DScore + Cutoff	0.52	0.63	0.66	0.66	0.66

Table 2. Precision@N with exact, and partial match criteria.

P@N (exact)	N = 1	N = 2	N = 3	N = 4	N=5
Glover’s	0.27	0.18	0.16	0.13	0.12
DScore	0.36	0.27	0.22	0.19	0.17
DScore + Cutoff	0.37	0.28	0.27	0.26	0.26
P@N (partial)					
Glover’s	0.39	0.32	0.30	0.28	0.25
DScore	0.53	0.45	0.40	0.38	0.35
DScore + Cutoff	0.52	0.46	0.43	0.43	0.43

Table 3. MRR, MTRR, and Average Length statistics.

Exact	MRR	MTRR	Avg. Length
Glover’s	0.35	0.38	5
DScore	0.47	0.53	5
DScore + Cutoff	0.45	0.47	2.6
Partial			
Glover’s	0.50	0.68	5
DScore	0.61	0.94	5
DScore + Cutoff	0.59	0.74	2.6

Both descriptive score models outperform the threshold-based approach. The Match@1 values are around 0.36 in exact match and 0.53 in partial match for both descriptive score models, compared to 0.27 and 0.39 for baseline model. This means that in almost half the categories, the descriptive score predicts the correct label with the top rank label. The precision of both descriptive score models is higher than the baseline model. This suggests that the lists of labels produced by our descriptive score contain more good labels than the ones produced by the baseline. This is also supported by the higher MTRR measure for the descriptive score model.

The average number of labels displayed with the cutoff model is 2.6. By choosing to display fewer labels, the algorithm with the cutoff model has a lower number of correct matches (M@N) and also lower MRR, and MTRR. However, the precision of the list of labels produced is higher, because the model tries not to show low-quality labels. We believe that the tradeoff in lower MRR with higher precision is worthwhile because a shorter list of labels makes it easier for users to understand the content of the cluster. However, a user study would be needed to verify our conjecture.

Figure 2. shows the labels produced by the DScore+Cutoff model along with the corresponding (“correct”) ODP labels. In most categories, the labels produced by the model match the category labels in the ODP. Even when model did not produce exactly the same labels as the ODP, the labels assigned by the model provide a similar description. For example, in the category, security / news and media, the list of labels, “attack, vulnerable, and hack” describes what most of the documents discuss.

The algorithm works well in spite of a very heuristic method used to generate scores for ODP labels during training. We

believe that this effectiveness is because the trained regression model does not need to predict an exact DScore; it needs only to produce a relative score for each label that is suitable for ranking them. One thing to note is that while the algorithm ranks labels using the relative importance of terms between the parent cluster and the self cluster, it does not use information about sibling clusters. The algorithm could potentially rank the same labels highly for multiple sibling clusters. However, in our evaluation with the ODP data, this was rarely the case. All sibling clusters are pooled together to form the parent cluster, so if the hierarchy is well-formed such that every sibling cluster is of roughly the same granularity, the highly ranked terms in the parent cluster are similar to the highly ranked terms of its children, yielding small relative differences. Comparing a child to its parent cluster has an indirect effect similar to comparing against its siblings. We suspect that in a less well-formed hierarchy the algorithm would need to consider information about each individual sibling in order to assign discriminative labels.

5.3.2 Using Only Rank Features

To test the hypothesis that one can identify a good label for a cluster based only on rank features, the descriptive score formula in Section 4.1 was modified to use only the rank features and the boost in ranking. Tables 4, 5, and 6 show the results.

Table 4. Match@N with exact, and partial match criteria for rank-features model.

Match@N (exact)	N = 1	N = 2	N = 3	N = 4	N=5
Glover's	0.27	0.35	0.42	0.46	0.50
DScore	0.35	0.52	0.57	0.59	0.64
DScore + Cutoff	0.35	0.52	0.55	0.55	0.55
Match@N (partial)					
Glover's	0.39	0.52	0.60	0.64	0.68
DScore	0.53	0.64	0.72	0.73	0.76
DScore + Cutoff	0.53	0.63	0.68	0.69	0.69

Table 5. Precision@N with exact, and partial match criteria for rank-features model.

P@N (exact)	N = 1	N = 2	N = 3	N = 4	N=5
Glover's	0.27	0.18	0.16	0.13	0.12
DScore	0.35	0.28	0.22	0.19	0.16
DScore + Cutoff	0.35	0.29	0.27	0.27	0.27
P@N (partial)					
Glover's	0.39	0.32	0.30	0.28	0.25
DScore	0.53	0.45	0.41	0.37	0.35
DScore + Cutoff	0.53	0.45	0.44	0.44	0.44

Table 6. MRR, MTRR, and Average Length statistics for rank-features model.

Exact	MRR	MTRR	Avg. Length
Glover's	0.35	0.38	5
DScore	0.47	0.53	5
DScore + Cutoff	0.44	0.48	2.5
Partial			
Glover's	0.50	0.68	5
DScore	0.62	0.94	5
DScore + Cutoff	0.60	0.77	2.5

The learned descriptive model based only on rank features is as followed:

$$\begin{aligned} DScore(p) = & 0.122 \\ & +0.0000 * r(DF_s / \# S) \\ & -0.0001 * r(DF_p / \# P) \\ & +0.0000 * r(TFIDF_s) \\ & -0.0001 * r(TFIDF_p) \\ & +0.0509 * [\log(r(DF_p / \# P)) - \log(r(DF_s / \# S))] \\ & +0.1874 * [\log(r(TFIDF_p)) - \log(r(TFIDF_s))] \end{aligned}$$

The model performs surprisingly well considering that it uses only ranking features. Its MRR is 0.47 for exact match and 0.62 for partial match definition, which are at the same level comparing to the models that use all features.

5.3.3 Noise Resistance

So far we have assumed that the document hierarchy given to the algorithm correctly clusters every document with the same concept together. However, this is rarely the case, because the hierarchical clusters that need automatic labeling are usually produced by imperfect hierarchical clustering algorithms. To evaluate how the algorithm performs in a more realistic setting, another experiment was conducted with noise introduced into our ground truth data.

Consider a cluster P that has the set of subclusters, denoted as $children(P)$. For each document in any subcluster of P , the document is reassigned to another subcluster of P with a probability N (Noise %); with the probability $1-N$, the document remains in the correct subcluster. The probability that a document is reassigned to a subcluster C of P is proportional to the size of the cluster C . So the probability that a document d in a the cluster P is assigned to a subcluster C of P , denoted by $Pr(assigned(d, C))$, is:

$$Pr(assigned(d, C)) = \begin{cases} 1 - N, & \text{if } d \in C \\ N * \frac{|C|}{\sum_{R \in children(P) \text{ and } d \notin R} |R|}, & \text{if } d \notin C \end{cases}$$

where $|C|$ denotes the number of documents that originally belonged to the cluster C .

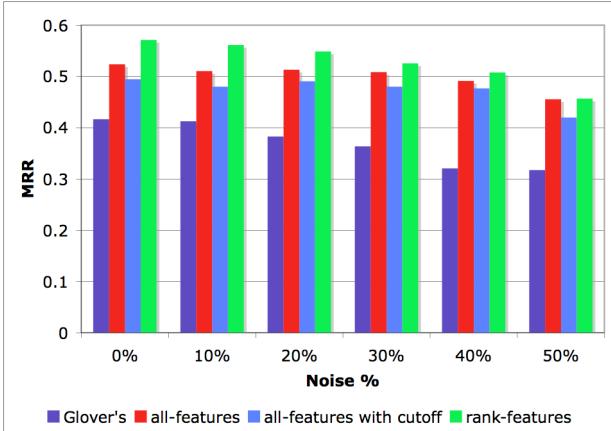


Figure 3. shows MRR for different noise probability levels

Figure 3 shows the performance comparison on exact match definition between different algorithms at noise levels from 0% to 50% on 94 categories of the OPD ground truth data. At 0% noise level, every document is correctly assigned. At 50% noise level, each cluster has approximately 50% of its documents correctly assigned, and the rest are documents that should be in its sibling clusters. We didn't investigate noise levels higher than 50% because the cluster identity is no longer coherent when most of the cluster is assigned incorrectly. In Figure 3, *Glover's* refers the baseline algorithm described in Section 5.3.1; *all-features* and *all-features with cutoff* refer to the models that calculate DScore based on every feature described in Section 4.3, with and without the cutoff model respectively; and *rank-features* refers to the model described in Section 5.3.2, which only used rank-related features to calculate DScore.

As expected the performance of every algorithm decreases as more noise is introduced. At 10% to 30% noise level, there is almost no change in the performance of any model. The MRR of every model drops around 0.1 with 50% noise. However, even with 40% noise level, our algorithms still perform at around 0.5 MRR, which means that on average there is a good descriptor in the top two labels. The decreases in performance mostly came from the small categories, which are more easily disrupted by the introduction of a few documents. In general, our algorithm was not sensitive to random noise.

One might argue that this result is not surprising, because randomly assigning documents to sibling clusters does not change the underlying distribution of words in each cluster. A more realistic simulation would assign a document to sibling clusters based on the similarities between the document and clusters' centroids. Such a simulation would better reflect the errors normally produced by a clustering algorithm. Although a further study is needed to assess the noise tolerance of the model under this scenario, we believe that our experiment shows promising initial result.

6. CONCLUSION AND FUTURE WORK

Tools that automatically organize and assist in the analysis of large amounts of text documents are becoming a requirement in many organizations. There has been considerable research on automatically organizing text documents into hierarchical clusters suitable for interactive browsing, but much less research

on how best to automatically describe or label hierarchies to support interactive browsing.

This paper presents a simple trainable algorithm that selects a few 1-3 word labels to describe each cluster in a document hierarchy. The algorithm dynamically decides how many labels to select for each cluster; in our experiments, it average about 2.6 labels per cluster. Experiments using Open Directory Project data demonstrated that the labels produced by the algorithm often match the labels chosen by human editors. Preliminary experiments suggest that the algorithm is also robust with respect to clustering errors, although additional research is required to settle this question.

Our research and most prior research focused on the use of statistical features to select and rank features; a distinguishing feature of our research is the use of statistics from a corpus of general English, the parent cluster, and the cluster to be labeled. However, perhaps more interesting is the discovery that the algorithm can select good descriptors using only rank-based features, and that rank-based features provide more robust results than more detailed numeric features.

Error analysis showed that most of errors come from clusters containing small numbers of documents. The small number of observations in small clusters can make good and bad labels indistinguishable; minor variations in vocabulary can also produce statistical features with spuriously high variance. To improve the performance of the algorithm on small clusters it may be necessary to incorporate lexical features, for example the number of word senses for a candidate label, or positional features sensitive to where terms occur in a document, for example in a title or in a lead sentence. The work described here demonstrates that it is realistic to aim higher than the lists of characteristic terms that have been the norm in prior research on automatic labeling, but it is nonetheless just the first step.

7. ACKNOWLEDGEMENTS

This research was supported by a Thai Ministry of Science, Technology and Environment Scholarship, and by NSF grants EIA-0327979 and IIS-0429102. Any opinions, findings, conclusions, or recommendation expressed in this paper are the authors', and do not necessarily reflect those of the sponsors.

8. REFERENCES

- [1] Caraballo, S. Automatic Acquisition of a hypernym-labeled noun hierarchy from text. In Proceedings of the Association for Computational Linguistics Conference, 1999.
- [2] Chuang S., and Chien L. A practical web-based approach to generating topic hierarchy for text segments. In Proceedings of the 20th International Conference on Information and Knowledge Management, 2004.
- [3] Cutting D. R., Karger D. R., and Pederson J. O. Constant interaction-time Scatter/Gather browsing of very large document collections. In Proceedings of International ACM Conference on Research and Development in Information Retrieval, 1993.
- [4] Glover, E., Pennock, D., Lawrence, S. and Krovetz, R. Inferring hierarchical descriptions. In Proceedings of the 20th International Conference on Information and Knowledge Management, 2002.

- [5] Glover, E., Tsoutsouliklis, K., Lawrence, S., Pennock, D., and Flake, G. Using web structure for classifying and describing web pages. In Proceedings of International Conference on World Wide Web, 2002.
- [6] Krovetz, R. Viewing morphology as an inference process. In Proceedings of International ACM Conference on Research and Development in Information Retrieval, 1993.
- [7] Lawrie, D., Croft, W. B., and Rosenberg, A. L. Finding topic words for hierarchical summarization. In Proceedings of international ACM conference on research and development in information retrieval, 2001.
- [8] Muller, A., Dorre, J., Gerstl, P., and Seiffert, R. The TaxGen framework: automating the generation of a taxonomy for a large document collection. In Proceedings of the 32nd Hawaii International Conference on System Science, 1999.
- [9] Pantel, P., and Ravichandran, D. Automatically labeling semantic classes. In Proceedings of the Human Language Technology and North American Chapter of the Association for Computational Linguistics Conference. 2004.
- [10] Popescul, A., and Ungar, L. Automatic labeling of document clusters. Unpublished manuscript, available at <http://citesear.nj.nec.com/popescu00automatic.html>, 2000.
- [11] Yang, H. and Callan, J. Near-duplicate detection for eRulemaking. In Proceedings of the National Conference on Digital Government Research (DG.02005), 2005.
- [12] Zeng, H., He, Q., Chen Z., Ma, W., and Ma J. Learning to cluster web search results. In Proceedings of International ACM Conference on Research and Development in Information Retrieval, 2004.
- [13] Open Directory Project (ODP).
- [14] eRulemaking Testbed. <http://hartford.lti.cs.cmu.edu/eRulemaking/Data.html>.
- [15] Weka, Data Mining Software, University of Waikato.
- [16] WordNet, a lexical database for the English language.
- [17] Yahoo!

Using Natural Language Processing to Improve eRulemaking

[Project Highlight]

Claire Cardie
Information Science Program
and Department of Computer
Science
Cornell University
Ithaca, NY USA
cardie@cs.cornell.edu

Cynthia Farina
Law School
Cornell University
Ithaca, NY USA
crf7cardie@cornell.edu

Thomas Bruce
Legal Information Institute
Cornell University
Ithaca, NY USA
trb2@cs.cornell.edu

ABSTRACT

This paper describes in brief Cornell's interdisciplinary eRulemaking project that was recently funded (December, 2005) by the National Science Foundation.

1. INTRODUCTION

Each year federal regulatory agencies issue more than 4,000 new rules [6]. By law, many of these must be created through a complex and expensive process in which the agency drafts a proposed rule and then exposes the proposal, any underlying data, and its legal and policy rationale to public comment. This process, *notice and comment (N&C) rulemaking*, is the mechanism through which most agencies make major regulatory policy. One of, if not the, most important functions of government agencies [6, 5]¹, N&C rulemaking is also one of the slowest. A duration of two to five years is not uncommon [5]².

In N&C rulemaking, the agency may receive anywhere from dozens, to hundreds of thousands, of comments, depending on the subject and complexity of the rule. The agency's fundamental legal obligation is to review all the comments received and, if it chooses to adopt the proposed rule, to issue a statement that not only (i) demonstrates why its choice is within its statutory authority and sound as a matter of regulatory policy, but also (ii) responds to significant criticisms made in the comments and explains why it rejected alternative approaches suggested there [10]³. The stakes for the agency are high. Failure to adequately address critical comments and discuss alternatives in the statement accom-

panying the final rule can lead a court to invalidate the rule thereby requiring still more agency time and effort to perform additional review and explanation.[10]⁴

The need to absorb and assess the significance of hundreds, or even thousands, of comments is not the only hurdle that confronts the agency trying to make regulatory policy through N&C rulemaking. Over the last 25 years, Congress and the President have imposed an increasing number of mandates on rulemaking regardless of regulatory subject area [9]. These mandates are typically designed to protect a specific interest (such as small businesses or Native American tribes) or are triggered when a proposed rule would pass a certain threshold (such as a certain dollar amount of economic impact). They may require that, before completing the rulemaking, the agency prepare a certain kind of analysis, consult with another agency or a particular private entity, or issue a specified certification. Rule writers have found it increasingly difficult to keep track of these mandates and to recognize which, if any, are relevant in a particular rulemaking [7, 9]. As a result, they may complete the long and expensive N&C process only to discover that an arcane but legally required assessment, consultation, or certification was triggered but not accomplished.

Electronic rulemaking (eRulemaking) includes a wide range of ways that information technology might be used in rulemaking. It includes, but is not limited to: converting the agency's docket (the filing system showing all its activities, including rulemaking) to electronic form and making it available via the Internet; allowing submission of comments via email and the Internet in addition to (and perhaps eventually in place of) conventional mail and fax; and using search engine, hypertext, and other IT capacities to allow both the public and agency rule writers to find, sort, and link the massive amount of material relevant in a rulemaking more easily and cheaply than could possibly be done with hard copies.

eRulemaking thus has the potential to radically transform the N&C process. It could make the process more transparent and accessible to the public, and more substantively

¹At 180,280-83.

²At 102-04.

³At 524-50.

⁴At 524-50 & 1016-26.

reliable and cost-effective for the agency.

To be sure, Module III of the eRulemaking Initiative contemplates developing a “rule writer’s tool kit” to help categorize comments, mine data, and provide online rulewriting instruction. While existing language processing techniques (e.g. for information retrieval, text categorization, document clustering, and information extraction) could provide some of the basic capabilities listed above, they would require significant testing and evaluation within the eRulemaking domain. In addition, research on methods that would clearly be invaluable in actually carrying out the more complex of these tasks has only barely begun. Work in the area of text summarization and sentiment analysis, for example, is still very new [8, 1, 2, 11], but will be essential to analyze and summarize the opinions expressed in comments.

2. PROJECT GOALS

Our propose to apply and develop a range of methods from the field of natural language processing (NLP) to create NLP tools to aid agency rule writers in:

- organization, analysis, and management of the sometimes overwhelming volume of comments, studies, and other supporting documents associated with a proposed rule; and
- analyzing proposed rules to flag possibly relevant mandates from the large number of statutes and Executive Orders that require studies, consultations, or certifications during rulemaking.

Officials from the Departments of Transportation and Commerce, with whom we are collaborating in the project, identified both tasks as high priority needs. We will focus on the use of information extraction, text categorization, and opinion-oriented text analysis techniques in both supervised and weakly supervised machine learning frameworks. Importantly, we will also focus on the use of human language technologies to elicit more informed comments from commenters. The tools and methods we develop should be valuable not only in the eRulemaking arena, but also in business (e.g. automatic analysis of online product reviews), government intelligence (e.g. analyzing emerging opinion on a hot topic in the Mideastern vs. European press), science (e.g. extracting information from biomedical literature to create a database), and social science (e.g. processing Weblogs).

We will evaluate the integration of the tools into the day-to-day rulemaking process by applying qualitative and quantitative methods from social sciences — survey instruments, longitudinal interviews, and direct observation [4].

More generally, we will study the effect of technology on the rulemaking process. Despite the crucial importance of rulemaking to federal regulatory policymaking, there is a serious shortage of research on how the process actually occurs within agencies [5, 3]⁵.

3. PLANS FOR 2006

Our plans for 2006 include a number of related efforts, each of which aims to proactively use technology, usually human

⁵Kerwin at 279-83.

language technology, to improve eRulemaking for rule writers and for the public:

- Begin the creation of an eRulemaker’s “best practices” guide.
- Investigate options for providing technical support for the creation of hyperlinks between (parts of) a proposed rule and relevant law.
- Develop ways to streamline the process of educating the public on the process and substance of rulemaking.
- Investigate options for employing NLP techniques to elicit better comments.

4. ACKNOWLEDGMENTS

This work is supported in part by NSF Grant IIS-0535099 and by a Xerox Foundation and a Google gift to the first author.

5. ADDITIONAL AUTHORS

Additional authors: Erica Wagner (School of Hotel Administration, Cornell University) email: elw32@cornell.edu.

6. REFERENCES

- [1] C. Cardie, J. Wiebe, T. Wilson, and D. Litman. Low-level annotations and summary representations of opinions for multi-perspective question answering. In M. Maybury, editor, *New Directions in Question Answering*. 2004.
- [2] Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan. Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns. In *HLT-EMNLP 2005*, 2005.
- [3] C. Coglianese. The state of rulemaking in the federal government. Technical report, Transcript Panel 6, 2005.
- [4] B. Kaplan and D. Duchon. Combining qualitative and quantitative methods in information systems research: A case study. *MIS Quarterly*, 12(4), 1988.
- [5] C. Kerwin. The state of rulemaking conference. Technical report, Transcript Panel 1 and 6, 2003.
- [6] C. Kerwin. The state of rulemaking in the federal government. Technical report, Transcript Panel 1, 2005.
- [7] T. O. McGarity. *The Expanded Debate Over the Future of the Regulatory State*. 63 U. Chi. L. Rev. 1463, 1523. 1996.
- [8] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In *EMNLP-2002*, pages 79–86, 2002.
- [9] M. Seidenfeld. *A Table of Requirements for Federal Administrative Rulemaking*. 27 Fla. S.U.L. Rev. 533, 535. 2000.
- [10] P. Strauss, T. Rakoff, and C. Farina. *Administrative Law*. 10th edition, 2003.
- [11] L. Zhou and E. Hovy. Digesting virtual “geek” culture: The summarization of technical internet relay chats. In *ACL-2005*, 2005.

SESSION 5C

TRANSPARENCY AND

E-GOVERNANCE

Moderator

Luis F. Luna Reyes, Universidad de las Americas, Mexico

Titles and Authors

Transitioning from e-Government to e-Governance in the Knowledge Society: The Role of the Legal Framework for Enabling the Process in the European Union's Countries
Paskaleva-Shapira, Krassimira

A National Center for Digital Government Program on Networked Governance
Fountain, Jane E.; Lazer, David

Connecting to Congress
Lazer, David; Esterling, Kevin; Neblo, Michael; Fountain, Jane; Mergel, Ines; Ziniel, Curt

Digital Deliberation: Searching and Deciding About How to Vote
Robertson, Scott P.

Transitioning from e-Government to e-Governance in the Knowledge Society: The Role of the Legal Framework for Enabling the Process in the European Union's Countries

Krassimira Paskaleva-Shapira

Karlsruhe Research Center
POB 3640, 76021 Karlsruhe, Germany
+49 (0) 7247 / 82 – 6133
paskaleva@itas.fzk.de

ABSTRACT

In modern European participatory democracies, government business and services not only rely on the trust of the business partners but moreover on the trust of the citizens. To be credible, transactions, services and interactions must be legally sound, provide value and be fair to all. In the Knowledge Society, traditional contract arrangements, in the context of e-Government, e-Business and e-Democracy are being redefined to reflect the complexity of the e-Governance ambiance, as the collective mindset changes towards citizen-centered service provision and institutional re-engineering. The legal frameworks, which underpin these changes, are undergoing reforms on international, state, regional and municipal levels. Continuous legislative change is considered essential to the success of e-Governance in context of the Knowledge Society and Sustainable Development. Yet, crafting new regulations imposes significant information and coordination demands among the agencies involved. This paper examines the role of law for European e-Governance, justifying the need of a cohesive legal framework as a common platform for effective electronic government ensuring democratic advance, transparent decision-making and inclusive policies in the local communities. It identifies key challenges that national and local governments face in coordinating and integrating European, national and regional legislation in the spirit of the main principles of Good Governance and Knowledge and Networked Society.

Categories and Subject Descriptors

K.5 [LEGAL ASPECTS OF COMPUTING]: K.5.2 Governmental Issues - *Regulation*

General Terms

Legal aspects

Keywords

E-Governance, Legal and regulatory framework, Public policy

1. INTRODUCTION

E-Governance is an ambitious challenge for a variety of stakeholders in the Information and Knowledge Society. Its potential, however, to modernize, transform and improve public management and policy is specifically acknowledged by the countries of the European Union. Its progressive advance depends, among other factors, on the legal and regulatory frameworks enabling its implementation across the European nations and localities.^[1] Identifying the legal aspects of e-Governance, the key issues and categories (such as ‘Legal Validity’, ‘Trust and Confidence’ or ‘Available Remedies’), the administrative structures, organs and players, the gaps, the barriers and the challenges to finding the best solutions is becoming a strategic pursuit in the Union.^[2] A cohesive Legal Framework of e-Governance, however, is just now starting to evolve. The current paper responds to this need and aims to support progress towards a better understanding of the legal aspects of e-Governance by identifying, analyzing, and assessing the specific legal aspects of e-Governance, the core laws and legal changes in the main areas that form the framework of e-Governance, the driving factors and conditions enabling the process on national and local levels, and the main challenges, key issues and barriers to the progress of transition from e-Government to e-Governance in the European Union. The key potentials of the Legal Framework as enabler of e-Governance are identified and endorsed with respect to four main legal areas: Personal data protection laws; Privacy and security laws; Administrative laws; and Information (provision) laws.

2. E-GOVERNANCE AND THE INFORMATION/KNOWLEDGE SOCIETY

2.1 The European Union’s Trust

E-Government¹ has emerged as a significant challenge in the Information and Knowledge Society, globally and specifically in the countries of the European Union (EU) where current policy trusts are comprehensive and focus on providing citizens centred

¹ E-Government has been mostly used for the services government agencies provide to the citizens using information technology. Despite emerging changes in its context and scope involving governance objectives and processes, the term continuous to be in use, yet with a new meaning.

effective services, improving doings of government, advancing the democratic processes and promoting inclusive decision-making processes for better public policies and management. Several strategic documents have outlined the process. In September 2003 the European Commission adopted a Communication, outlining the importance of e-Government as ‘a means of achieving world-class public administration in Europe’ [3]. Earlier, in 2000 the EU’s Lisbon Strategy reinforced e-Government as a ‘potential provider of major economic boost by facilitating new and better services for all citizens and companies’ [4]. Equally important, e-Government is being granted the task to further ‘reinforce democratic development in Europe’ [5]. These new missions of e-Government provide a turning point for transitioning to e-Governance where governing is inclusive of the other stakeholders of the civil society, participating and/or impacted by its processes and activities. Implementing e-Governance on various levels however requires a cohesive legal framework that will facilitate the meeting of its objectives aimed overall to generate a new public value for each of the European nations and the numerous local communities. Yet, discussing the legal aspects and challenges of e-Governance requires setting up the concept of e-Governance in its European contexts along with the opportunities, the barriers, and the challenges involved.

2.2 Definition of E-Governance

The case for e-Governance has been already effectively made by a large range of international and regional stakeholders, including the United Nations, the European Union, the OECD, the World Bank, at the countries level, by private organizations and academic institutions. As a result, the existing studies and findings on the issues are abundant. Similarly are the views, the approaches, the definitions, and the legal aspects considered? Because of its multi-dimensional aspects, e-Governance poses serious challenges to conceptualizations. Many organizations and players have already attempted to define it based on their approaches. Some, like the OECD, centre on its functions (functional definitions), others focus on the processes (descriptive definitions), third (such as the European Union) refer to its essence (conceptual definitions). Few define e-Governance in reference to e-Business, while the World Bank, for example, combine multiple elements and offer a rather complex definition [6]. In spite of the different approaches and perspectives, however, there is often a convergence in the approaches and the definitions. A common consensus has emerged on the challenges of e-Governance for the Information society along with some key words - ‘public administration’, ‘efficiency’, ‘information infrastructure’, ‘quality service’, ‘governance’, etc. Establishing the right definition of e-Governance is a pre-requisite for any serious work, analysis and decision that carry legal consequences.

The European Commission defines e-Government as: “*The use of ICT in public administrations combined with organisational change and new skills in order to improve public services and democratic processes and strengthen support to public policies.*” [7] The definition refers to the fundamental essence of e-Governance as ICT-based applications, government-related actions, and government role in steering participation and societal advance. The main idea behind this definition, however, is that e-Governance is more about ‘Government’ than ‘e’, a reforming

and innovative government involving citizens and businesses in the decision-making processes and acts. In the next sections we attempt to define the specific elements of e-Governance, as in the above context, identify their legal aspects and areas of concern, and map out the challenges in establishing a cohesive legal and regulatory framework of e-Governance in line with EU’s strategic objectives.

2.3 Europe’s e-Governance Policy Agenda

Within both central and local governments in Europe there is a strategic agenda to radically transform the delivery of public services through the adoption of advanced ICTs. Likewise, as information technologies increasingly penetrate the public sphere, governments contemplate the use of these tools to remodel democratic practice and transform relations between citizens and public services. E-Government now becomes an issue of not solely technology – it is also about reinventing the way in which service providers and customers interact and transform government processes, providing leadership, enable economic development, and reinvent the role of government itself in society [8]. Integrated public services and an innovative organizational change in government to citizen relationships, including citizen-centric services, participation and open interoperable frameworks are therefore necessary to address the challenges of the Networked Knowledge society. The EU seeks a high level of integration of policies and laws. As far as e-Governance is concerned, a range of initiatives have driven the process. In 1999, the “eEurope: An Information Society for All” Initiative set an ambitious objective to ‘bring the benefits of the Information Society within the reach of all Europeans in an inclusive, trustworthy and secure fashion’. Promoting Digital Government that ensures all citizens’ easy access to government information, services and decision-making procedures is now a key trust of the Commission, the Member States, industry and the European citizens in the Knowledge Society [9]. The eEurope 2005 Action Plan and the 2003 Commission’s Communication on ‘e-Government for Europe’s future’ further affirm ‘facilitating modern on-line public services’ is a key European strategic challenge. The 2004 eEurope Advisory Group Recommendations on e-Government Beyond 2005 reinforces the need of progress in implementation - ‘e-Government is the way forward and a catalyst for innovation, therefore should now deliver its promise’ [10]. The challenges, however, are diverse and comprehensive. Identifying the true essence of e-Governance, setting up a strategic e-Governance structure, promoting regulatory and policy innovation and coordination, seeking cooperation in implementation, transformation and finance are some of the issues that still need to be addressed.

2.4 Elements of e-Governance

To establish the main legal issues raised by e-Governance and the barriers to their implementation, the focus of this paper, requires identifying the main aspects and objectives of e-Governance with view of the European normative, value and policy systems. To begin with, it is important to set up the framework of e-Governance. A general parallel with e-Commerce, for example,

might be necessary. Indeed, the paradigms are fundamentally different. While e-Commerce is business-driven, e-Governance is, or should be citizens or people centred.² Efficiency and quality services, however, are too fundamental to e-Governance. The core idea of ‘public service’ or ‘public administration’ is therefore central to delivering successful e-Governance. The 2003 Commission’s Communication on e-Government makes a particularly important statement with regard to the core role of Public Administration:

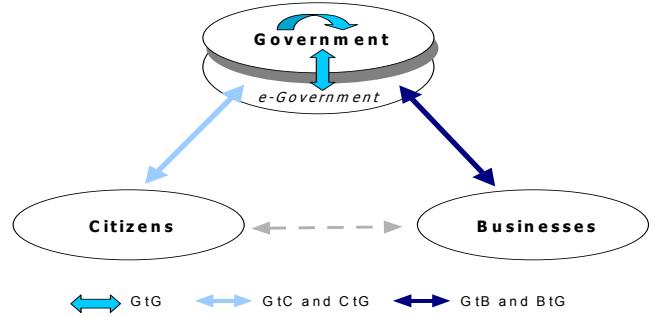
“Within the public sector, public administrations are facing the challenge of improving the efficiency, productivity and quality of their services. All these challenges, however, have to be made with unchanged or even reduced budgets... ICTs can help administrations to cope with the many challenges/ however the focus should not be on the ICTs itself. Instead, it should be on the use of ICT combined with organizational change, and new skills in order to improve public services. Democratic processes and public policies. This is what e-Government is about”.

Clearly, e-Governance is not simply the process of moving existing government functions to an electronic platform. Rather, e-Governance is more about ‘government transformation and modernization’ than just ‘e’. It is also about ‘democratic processes’ and ‘public policies’. E-Governance reflects the capacity and ability of government to reform and improve to better serve its citizens. It also means engaging with the stakeholders to share the risks, opportunities and benefits of collaboration in steering the nation’s and community affairs [11]. It offers potential solutions to leaders to better assume their responsibilities. Finally, it invites people to participate in making the decisions of the nations and the local communities. With respect to policies, e-Governance offers a tool for government agencies to facilitate effective decision making and improved public policies by transforming relations with citizens, businesses, and other arms of government. To fulfil its mission, however, e-Governance has to reflect the principles and objectives of ‘Good Governance’, which include government efficiency, transparency, openness, accountability, and inclusiveness.

2.5 Relationships and Interactions

From the user perspective, within the area of Public Administrations, e-Governance has three main components: Government to citizens (G2C); Government to government (G2G); and Government to businesses (G2B). In all cases, the relationship is two-fold between the two parties with the government steering the cooperation (Figure 1).

Figure 1: e-Governance and Stakeholder Interactions



2.6 Obstacles and Challenges

Despite the merits, e-Governance can be hampered by a variety of impediments, among which are legislative and regulatory barriers, policy inconsistencies, budgetary frameworks, digital divide, institutional traditionalism, bureaucracy and lack of action. Successful implementation is said to require addressing a set of key challenges: Reorganizing government organizational capacity; Strengthening ICT structural capacity; Strengthening stakeholder participation capacities; Positioning the nation or the locality in context of the EU, the World [12]. Leadership and long-term commitment are necessary from policy-decision-makers, public administrations senior managers and companies to drive the process. Yet, the legal framework is considered critical to effective implementation and to the success of e-Governance in the Knowledge society [13]. World-wide, the large-scale use of e-Governance has been attributed primarily to ‘legislative professionalism and professional networks’ as a recent study that evaluated trends in the adoption of e-Government across the USA conclude [14]. By and large, promoting legal changes hand in hand with service and technology innovation and government reform is considered essential to the success of e-Governance across nations and municipal communities.

3. E-GOVERNANCE LEGAL ASPECTS

3.1 The Enabling Framework

Implementing e-Governance, as OECD recently confirms, ‘can be risky, expensive and difficult’ [15]. To get it right, necessary is to establish its legitimacy, secure the efficiency and the trust between government and citizens, and the legal aspects become part of its roadmap, strategies, and long-term objectives. Yet, the e-Governance legal frameworks are still in infancy [16]. A cohesive legal framework may help facilitate a better progress. The present section draws attention on some of the legal challenges of e-Governance aiming to promote a further discussion on the issues and the potential solutions involved. Success of e-Governance in the EU is strongly attributed to the institutional, legal and regulatory structures that underpin it [17]. Back office integration and cooperation based on clear regulatory frameworks is considered critical to the process. User confidence and trust in government electronic information, transaction and interactions is another ‘must’ prerequisite [18]. Yet, legal and regulatory frameworks can be both enabler and/or impediment to e-Governance. The European Commission’s e-Governance

² An UN based principle approach supported by the OECD, Rand Europe and other world organizations. For more information on the discussion see Paskaleva, K. 2004. ‘The regulatory conditions and legal framework as an e-Governance enabler’, in Ásta Þorleifsdóttir (et al), ‘City e-Governance: Best Practice Report’, EU IST 6FP IP INTELCITIES Report, Deliverable 11.2.2, www.intelcitiesproject.com.

approach highlights the political and policy dimension of e-Governance which implies institutional and legal reforms. The key words used indicate the areas of concern of the Legal Aspects of e-Governance - ‘better public services’, ‘efficient delivery’ ‘enhanced democracy’, ‘improved policies’, ‘better government’, ‘effective government management’, and ‘public participation’. Both, the role of Government and the impacts of e-Governance on Public Administrations (and Society) from the perspectives of legitimacy, transparency, accountability, integrity, efficiency, participation and the rule of law are inherent. Adding ICTs to Government is not sufficient to reach these objectives. All layers of e-Governance must undergo fundamental transformations, including the legal aspects. As noted, e-Governance is more about Government than ‘e’. Identifying its legal aspects therefore requires a good understanding of Government. Political science defines ‘Government’ as a series of Acts of political, legislative, administrative, and authoritative nature performed by the State, State entities or Administrations which affect citizens’ life [19]. Government is also about ‘Governance’, i.e. the interactive and participatory process by which government is exercised and shared by a public authority in the country. Government Acts include, among others, Law making, Justice, and Provision of services, Administration, Health care, Education, Utility Management, Planning, and Transportation. Traditionally processed on paper, with the help of the ICTs citizens can access, receive, contest or claim Government Acts electronically. E-Governance provides the possibility to establish a more open, inclusive, productive and trustworthy public sector. In the context of e-Governance, ‘Good Governance’ can be achieved by adequate combination of information technologies, organizational innovation and improved capacities and skills.

3.2 European e-Governance Regulation

E-governance and, in general, the trend towards general use of ICT requires adequate regulation of aspects such as human rights protection with regard to the processing of personal information, protection against ICT crime, ICT security, probative value of electronic information, electronic signatures, equal access to public services, transparency of administration, etc. In a globalized world, where many processes aren’t confined to one country’s territory, many of these rules need to be co-ordinated internationally. However, each country has an obligation to implement the principles established internationally into its own legal system. The EU has been much active in working out co-ordinated rules in a set of domains mentioned above. Generally, this is done by promulgating directives. The following directives are relevant in this respect:

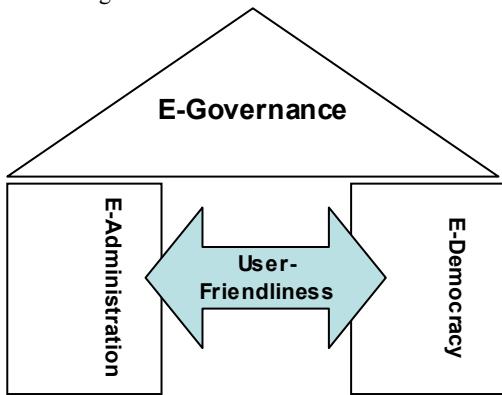
- Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data [20];
- Directive 97/66/EC of the European Parliament and Council of 15 December 1997 concerning the processing of personal data and the protection of privacy in the telecommunications sector [21];
- Directive 1999/93/EC of the European Parliament and of the Council of 13 December 1999 on a Community framework for electronic signatures[22];
- Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market (Directive on electronic commerce) [23];
- Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications) [24].

4. GOVERNANCE ISSUES

4.1 Synergy of e-Administration and e-Democracy and their Legal Systems

In line with the ‘Good Governance’ principles, e-Governance combines elements of both e-Administration and e-Democracy linked by key values of e-Business. (Figure 2). In the current European legal discourse, three main sets of legal issues are being considered as relevant to e-Governance: Legal validity, Trust and confidence, and Available remedies [25]. In regard to the *e-Governance Legal Validity*, core is the process of dematerialization of government activities by electronic means and open network communication and migration of paper based documents to electronic documents. Examples of legal validity issues are Law enforcement and Establishment of rights and titles. Adoption of laws recognizing e-Governance is the first step, however. Defining the e-Governance main principles, the objectives and the responsibilities of Government to achieving these objectives is the central point. Key areas for concern in the domain include document handling, data mining, and use of electronic signatures. With respect to *Trust and Confidence*, it is clear that public services can be delivered only in an environment where trust and confidence are secured and sustainable. The requirements concerned refer to the protection of personal data, identity management, authentication, privacy, network and information security and the fight against cyber crime. The range of issues to be addressed is here wide. Most of them are already covered by provisions of current Laws and regulations in the countries of the EU. Yet, practices indicate multiple problems of implementation, discussed earlier in this paper. Finally, how much *Remedies* does e-Governance offer to the citizens? Or does it observe the rights and obligations of citizens and governments endorsed by International and European Conventions - is another area of legal concerns. The specific legal aspects of e-Governance here refer to the process and interaction, both fundamental to ICTs based applications and government actions. Processes are fundamental to defining obligations, protect legal rights and determine liabilities. The legal issues to be addressed refer to the remedies available to citizens in e-Governance. Establishing and providing access to e-State Justice Systems and creating alternative to State Justice such as Online Dispute Resolution Mechanism have been acknowledged as central to providing effective remedies to citizens to seek expeditious and fair justice against the liabilities of the administrator [26].

Figure 2. Elements of e-Governance



4.2 The E-Governance Legal Framework

The regulatory and legal framework of e-Governance is broad and complex, due to the wide variety of laws and regulatory agencies involved. Another element of complexity derives from the variations in the ways that national regulatory systems are organized in different countries, with varying mixes in terms of public and/or private provision and utilization, and whether or not general practitioners play a gatekeeper role in determining access to other services. There are significant variations in the ways that services are delivered and in what is deemed acceptable or good practice. For this paper, the e-Governance Legal Framework is defined as: '*The combined set of laws, decrees, regulations and programs which facilitate e-Governance implementation in the spirit of the Good Governance principles and objectives.*'

The main domains of concerns are therefore defined as: Back office re-organization; Inclusive access; Trust and confidence; and Better use of public sector information. Four main sets of legislation are considered relevant: Personal data protection laws; Privacy and security laws; Information (provision) laws; Administrative laws

Regarding interactions and participation, the legislative issues of concern refer to three main schemes of relationships identified earlier (G2G, G2C, G2B). As argued earlier, the regulatory issues of e-Governance cross cut areas of the e-Governance, e-Democracy and e-Business legislative frameworks. Continuous and coordinated change in all three areas is therefore essential to facilitate the establishment of an effective e-Governance legal framework. The latter can provide an opportunity to re-think the whole process of government organization in back and front office and a reorganization of the whole public service supply chain [27].

4.3 Levels of Legal Compliance

The legal aspects of e-Governance in Europe can be referred to several layers of compliance: International frameworks and guidelines (UN, OECD); European Union's laws and regulations; National and/or regional legislative frameworks; and Local policies and regulations providing implementation. The legislative frameworks relevant to e-Governance are, however, generally a national priority. They provide the legal grounds for national,

regional and local policies relevant to ICT and government organisation.

4.3.1 European level

Increasing participation in democratic activities is a priority governmental issue in the countries of the EU; therefore developing tools for this purpose is to be bound by a legal framework. It is the responsibility of all governments to ensure that products of the information society delivered electronically are products for all people. This responsibility is based on ethical obligations as well as on national, European and international legislation. General codes can be found in international human rights agreements and in various recommendations on equality, entailing non-discrimination. The UN's Universal Declaration of Human Rights [28] recognizes that 'all human beings are born free and equal in dignity and rights'. There are also relevant, non-binding documents, such as the United Nations' Standard Rules on the Equalization of Opportunities for Persons with Disability and the EU's Amsterdam Treaty prohibiting discrimination, for example, on the grounds of disability [29]. Article 8 of the European Council's Convention for the protection of human rights and fundamental freedoms affords the right to 'respect for private and family life to every EU citizen'. In the European security and privacy measures, integration of national and international laws are observed – notably the OECD Privacy Guidelines [29], the new European Electronic Regulatory Framework, and the National Acts. The EU member states have adopted various regulations impacting the development of e-government. The analysis of main laws and acts in several countries highlights common tendencies but also national specifics in building national legal frameworks for action. Most EU member states have adopted legal texts covering key themes such as freedom of information, access to public information, use of public sector information, data protection and privacy, e-Commerce, Communications and e-Signatures. Through the electronic signatures Directive and the data protection legislation, the EU provides the rules to secure electronic communications. The privacy Directive constitutes a vital element in the new 2003 European Electronic Regulatory Framework [30]. It is designed to protect the fundamental interest of the end-users' and requires national transposition measures of draft laws. In the Community legislation, 'access' is a generic concept covering all forms of access to publicly available networks and services, whereas 'interconnection' refers to the physical and logical linking of networks. The Commission recognises the fundamental importance of the provision of access and interconnection services.

4.3.2 National level

Cultural, institutional, and legal conditions influence success of e-Governance in the states of the EU [31]. National Legal Frameworks include both general and specific elements, along with the administrative bodies involved. Some countries, such as France, have specifically defined the role of e-Governance on line in their Information Technology Programmes and e-Administration is boosted by a series of measures. Austria has adopted its first e-Government law in March 1, 2005 which is expected to greatly enable the introduction of a new Smart Card

[32]. Laws and regulations are often a result of national debates and consultations, as in the case of France where as early as 1960s issues as access to information and diffusion were put forward for public discussion, particularly with regard to the rights and protection of citizens in regard to their relations with Government. E-Business legislation often covers e-Governance activities too. A web of administrative structures and organs at the national and local levels enforce e-Governance policies. Some of them are related to sectoral ministries, others are directly linked to the Cabinet or the Prime Ministers Office. Main players include e-Governance agencies, Committees on Information Society Services, or Inter-ministerial committees on territorial access to information technologies or modernization of public administration. A number of Member States are undertaking legislative changes to facilitate e-Governance enhancement, among which leading are the Netherlands, Spain, Portugal, and Finland [33].

5. LEGAL CONSTRAINTS AND SOLUTIONS TO E-GOVERNANCE

E-Governance implementation in the EU countries depends mostly on national and local legal frameworks. Understanding their specifics and the potentials is central to its implementation.

5.1 The National Experience

A January 2004 EU Report [34] on ‘European good practices in reorganising government back-offices for better electronic public services’ affirms there is a clear and strong link between reorganizing government back-offices and the electronic public services available to the people, in other words, between the front and the back-office. Centralization of back-office and /or of data sources and decentralization of front-office functions is an important strategy in e-Governance driven by the need to increase efficiency and provide more effective, higher quality services. Yet, legal (along with political or legacy) barriers may hinder these processes. Legal restrictions, for example on data sharing between agencies, can pose difficulties to data integration between administrative and regional entities. Yet, countries like Austria, Spain and Denmark have shown that this could be done on various levels if there is indeed a sufficient will to do so. Federal structures and multiple levels of government can also impede the process of integration (for example Germany). Laws and regulations relating to government interactions with citizens (such as the legal validity of the interactions between the citizens and the agencies, or the amount of information on citizen that a public office is allowed to integrate in their dealing with the citizen) can be major obstacles for services that are feasible and logical from technical, organisational, and citizen’s point of view. Administrative laws too often allow institutional and cultural frameworks which prevent government back-office integration and centralization. Indeed, setting up adequate legal and regulatory conditions for back office reorganization have bared some positive results. In the ‘Bremen Personal Documents’ case in Germany for sharing of birth and marriage certificates a close cooperation was established between the city (a small ‘city-state’) and the Bremen federal authorities to overcome

complexities of different levels of e-Governance jurisdiction (federal and local government). The ‘Citizens Portal’ in Denmark, on the other hand, is experimenting with a system to enable the citizens to directly access and combine their own data located in different sector agencies, though the latter themselves are legally barred from doing so because of data privacy restrictions – this despite the fact that existing laws and legislation relating to interactions between the public and citizens practically hinder services that are feasible and logical from an organisational, technical, and citizens’ perspective. All member states have transposed the EU Community framework for *electronic signatures* (1999/93/EC) between 1997 and 2003; most countries have also transposed the European Directive 2000/31/CE on Electronic Commerce, except Greece, Slovakia and the Netherlands. Good practice example are the Spanish ‘Citizen’s Income Tax’ case, where legal changes opened the opportunity for on-line services overcoming problems of identification, authentication, and privacy. These changes have accompanied the development of the service and have led to an actual tax declaration system proving the possibility of an intermediary presenting the declaration on behalf of the tax payer. In the ‘Esslingen Building Permission’ case of Germany a change in the regional law was necessary to legally bind building plans on-line without requiring equivalent paper documentation. This has increased on-line applications and reduced administrative costs.

Public access to government information is considered an essential feature of advanced democracies [35]. In Europe, this principle is strongly endorsed, for example by the Dutch Government in regard to digital public service. Yet, openness to government information and transparency is not equally advanced all across the continent – only a few countries like Sweden, Denmark, Finland, Ireland and Portugal have legislation regulating access. In others, for example Germany, it is entirely unregulated.

The *protection of personal data* is at the centre of the trust question in e-Governance [36]. The Security and Privacy Laws (relevant to security of information, personal information protection, identity authentication, public safety, etc.) play a key role in this. In the context of e-Governance the main challenge is to ensure a balance between these two areas and create a climate of trust. Key challenges here are: To balance security, privacy and civil liberties, including those arising from the economic (e-commerce) dimension; The growing issues of identity theft and identity fraud; The state of the public opinion and; The concerns that arise from the privacy perspective in particular [37]. In the EU countries, texts on data protection and privacy were mostly adopted earlier than dispositions on freedom of information, except in countries such as Austria, Sweden, the Netherlands, Portugal, Italy, Denmark, the Czech Republic or Slovakia. In these countries, freedom of access to public information has constituted the basis on which to build the legal framework when in other countries, the framework was to a larger extent built on personal data protection and privacy laws. Personal data protection and privacy texts exist in all states. However, most countries adopted the texts after 1995, in application of the EU’s Data Protection Directive (95/46/EC), except in Sweden (1973), France (1978), Hungary (1992) and Belgium (1992). In such

countries, as well as in Greece, no specific texts in relation to freedom of information were adopted but data protection was implemented for a longer time than other member states [38]. Existing legal frameworks, however, in many cases, are reported to impede e-Governance – a recent UK survey finds ‘E-governance progress is in danger of being held back by the current legal quagmire of data protection and freedom of information law [39]. Underlying many security-related technology applications is the issue of ‘data mining’. New technologies can facilitate collection, cross-referencing and sharing of enormous amounts of information. While this can be of considerable benefit to the security and law enforcement fields, regulations concerning the nature, purpose, and access to such information should be expanded and made more specific to protect the privacy of the citizens [40]. An important element is to lay out the ground rules for longer-term data retention and re-use. It is this latter element in particular, that provides the most pervasive point of intersection between security and privacy issues – and ultimately the most fundamental tension [41]. Yet, countries such as Canada [42] and Hong Kong [43] have demonstrated that legal changes can facilitate a balance between the privacy and security considerations in e-Governance. In Europe, in the Netherlands, legal work has been carried out to ensure that the fundamental rights of citizens are protected in the digital age [44]. In Iceland, the legal framework allows the citizens to make use of emails for formal communication with the government [45].

Providing information on its activities is a major responsibility to government in liberal democracies, relevant to the principles of open and accountable governing. ICTs have given public authorities an unprecedented power to achieve this objective effectively and on a massive scale. Across the world and Europe, however, practices vary in context of the amount and nature of information which governments opt to provide and share with its citizens. In some countries (for example Greece, Ireland, Hungary, Denmark or Italy), freedom of information and use of public sector information aspects are embedded in the same legal text for a long time, especially in Denmark (1985) and Italy (1990). In most countries however, there is no specific law or regulation regarding the use of public information, except in Lithuania (2000) and Poland (2001). A ‘good practice’ example of openness and inclusiveness is the national government of The Netherlands – an amendment to the Constitution is proposed to add a new fundamental right – a personal right to all members of the public to have access to government public information, an action strongly supported by Government [45]. A raft of legislation in Belgium resulting from the Copernicus Reform has supported the introduction of more transparent government with simplified access for citizens.

5.2 Local level Practices

National and EU relevant legislation for e-Governance set up the broad limits within which localities operate. The latter, however, hold the powers to promote local regulations and implementation guidelines of either national or regional legislation regulating IT structures and access, providing there is a will to do so. They may also formulate legislation regarding local ICTs content which can

advance governance processes in the community. Good practices in this regard are rare but examples are becoming more and more available. The introduction of e-Governance and a well-planned modernization programme to integrate and deliver the range of services available to citizens has changed the face of Catalan local government. The result is a simplified process for citizens and local businesses to access a wide range of information and services available within all tiers of government which are delivered through a unique interface [46]. In a recent ‘City e-Government European Survey’ of a current IntelCities IST Integrated project, part of which is the study presented in this paper, administered in 11 European cities in Europe [48], when asked about existing legislative frameworks supportive of urban e-Governance, majority of the city officials indicate a general lack of a true e-Governance context applicable to their communities in the emerging acts on the national level. Among the variety of issues raised, the problems encountered in facilitating communication with the citizens, for example, and the implementation of digital signatures for the city administrative processes, were scored as a major legal issue, that is yet to be resolved. To the question, however, how important is the e-Governance legal framework in regard to impacts in the city, ‘open government’ and ‘better local policies’ were singled out as the primary beneficiaries, while ‘advanced democracy’ and ‘transparent and effective decision-making’ were generally on the ‘losers’ side. All city officials state, however, that although technology is an effective mean to provide electronic services to local citizens, there is still a need for a concrete and improved legislative framework in e-Governance.

6. LEGAL BARRIERS AND CHALLENGES TO E-GOVERNANCE IN THE EU

The issues here are as important as they are diverse. Here are some of the most pressing ones, split in four main categories which need further attention:

6.1 Defining the e-Governance Framework

Defining e-Governance is not an easy task. Different approaches of organizations, national laws and institutions have given e-Governance different meaning. Whether it’s the functions, the essence, the concept, or the processes that sit in the centre, however, the main idea should be that e-Governance is not just about ICTs. There should be a clear understanding about the importance of e-Governance for Government its self and the public it serves. Efforts should aim objectives and actions that best fit and further enhance existing structures and capacities. Setting up the conceptual and implementation framework of e-Governance by Government and citizens is the first step to successful e-Governance. External and internal factors are also to be considered. Important questions, yet to be addressed include: Which are the main factors that impede Governments to deal with e-Governance in a cohesive and people-centred framework? Who are the main players in this scheme? How can different levels of government work together to set up a common agenda that serves all citizens? What factors can promote better relationships between different levels of government that facilitate a common

agenda and strategies? What further policies must be promoted to support implementation?

6.2 Establishing and Enabling the Legal and Regulatory Foundation

One of the most evident conclusions emerging from the present discussion is that the legal and regulatory factors can be extremely important in determining the nature, the effectiveness and the success of e-Governance. Continuous improvement and coordination of e-Governance legislature is necessary to support implementation of the process. Much more importantly, there is sufficient evidence today that e-Governance relies on the development of a clear and comprehensive legal framework spanning across e-Administration, e-Business and e-Democracy which helps government improve its performance continuously and build public confidence and participation. Progress clearly requires consideration of both organizational and technological conditions which often reflect national and/or regulatory and institutional regimes and cultures. Promoting legal changes hand in hand with service and technology development is essential to the success. Cohesive legislative and regulatory frameworks are necessary to bolster government effectiveness in improving governing, advancing democracy and fostering transparent decision-making and inclusive policies in the nations and the local communities. The challenge is to promote legal changes and regulatory conditions, which provide for back office modernization, inclusive access, trust and confidence, and better public sector information to all. This, however, requires synchronized action in the administrative, data protection, privacy and security, and information provision laws involved. Coordination between the European, national and regional legislation is necessary to allow public administrations develop effective electronic services and policies in the spirit of the main principles of Good Governance. The challenge, however, is to encourage government to foster the regulations and the mechanisms ensuring an ICT infrastructure and Good Governance mechanisms that would best serve their citizens and the local communities. For this to succeed there needs to be the will to do so. These are some of the main challenges:

- *Legal changes facilitating fundamental human rights in the digital age.* Certain questions here need to be addressed: Are our fundamental rights sufficiently ICT-proof? Are these fundamental rights formulated in national constitutions, giving the legislature a firm enough basis in the Information Society? Do they give the public sufficient protection? Does development in ICT mean we need to develop new fundamental rights? Are the existing legal frameworks in general applicable to the Information Super highway, i.e. one set of standards for on-line and offline communications? The key point being here that fundamental rights have to be able to stand the test of time [49].
- *Legal framework acknowledging the right to informational self-determination* with the key question to be addressed being ‘Are national constitutions to provide for digital communications between Government and the public to be subject to the same principles as those governing other areas

of public life, including the general right to privacy of citizens?’

- *Providing access to all citizens* where nations have to decide whether to incorporate rights of access to government information in their constitutions. Other important questions include: Do governments act accordingly? How much easy must access be to allow use and participation? How much information must Government make available to the public? Are current legislation and court rulings available online free to the public? A core point here is the link of access to the principle of inclusive government and public involvement in decision-making and policy.
- *Protection of personal information* that has to respond to the core issues: Do national and regional legislature on data protection ensure the provision of the key principles of good information management, such as accountability; identifying purposes; consent; limiting collection; limiting uses; disclosure and retention; accuracy; safeguard; openness and individual access? Does current legislation provide the best for both good security and good privacy protection? [50] Is privacy legislation compatible with the international and national Charters of Rights and Freedoms [51], which interpret personal privacy in the context of the fundamental elements of Western democracy and therefore represent the greatest single safeguard to adoption of overly intrusive security legislature? Are human behaviour and knowledge of threats and remedies considered in privacy legislation? Are information security concerns considered in conjunction with other policy fields such as individual privacy, industrial policy, international trade, citizens’ rights, law enforcement, defence? Are holistic approaches at both European and global levels taken into a consideration in national Laws? [52] Are security and privacy issues addressed in conjunction – to find the right catalyst so as the final solution is a hybrid – dominated by neither privacy nor security issues? [53] Are institutional safeguards in place to make privacy meaningful? Do governments include in their legal definitions of privacy to account for the changes that ICTs have brought to the relationships between transparency and confidentiality? [54]

Clearly, an adequate consideration of these issues in the relevant laws and principles can assist the process of e-Governance in the EU. National as well as federal and local initiatives are here necessary. The latter in particular, being close to their social and business communities, can strike a balance between the diverse interests and the local autonomy and freedom responsive of the specific local needs. Different localities may develop their own paths reflective of their unique identities, legal and institutional structures and cultures while complying with national and international standards to ensure integration and interoperability.

6.3 Undertaking Organizational Changes

E-Governance progress often depends on whether or not there is a will and capacity of public administration to promote the regulations and framework to implement e-Governance. If there is one, it is relatively quicker and easier to achieve significant benefits of e-Governance. If not, this could be a relatively long

and complicated process to manage and coordinate change. Implementation of e-Governance must be based on cohesive e-Governance policies, based on the specific needs and the long-term integrated strategies in ICT and governance issues. While availability of funding and training capacities is core to the success of the former, legal validity, trust and confidence and available remedies are fundamental to the latter. A principle point of consideration is the coordination and complimentarity of the different policy levels - EU, national, regional, and local – which together should provide standardization, harmonisation, integration, communication and participation in e-Governance. Most important, however, is the acknowledgement that a successful e-Governance strategy is based on of a strong political will, dedicated and capable institutions, and participation of the citizens and the civil society. The legality and e-Governance is a main starting point. A number of challenges are to be addressed here: How can public administrations make a radical shift to network, face integration and enhance local democracy in our nations and cities using the new technologies? What are the key conditions, policies, events and arenas of change? How can citizens and businesses have real benefits from on-line public services and participation? Where are we today in reengineering the government processes? How can e-Government be transformed to convey to the core principles of Good Governance? How can policies across the EU, nations and localities, highlight the key challenges of these topics, additional work is necessary in context of the legal frameworks involved?

6.3 Learning from Best Practices

Finally, in order to deepen and strengthen the eEurope approach for leveraging Good practices a reinforcement of exchange of these in e-Governance is necessary.[55] Good practices must encompass the technological, organisational, legal and training elements of successful e-Governance, this requires long-term commitment of all key actors involved and practice must illustrate tangible benefits and results. While demonstrating the state of the art, best practices must also point to new requirements for regulatory frameworks, change management, and organisation of work within administrations. They must help to identify research challenges and form a contribution to establishing a European Research Area in e-Governance [56]. With view of the legal aspects, the specific challenges refer to the core laws and legal changes in the key areas that form the framework of Good Practice e-Governance; the key characteristics of existing regulatory frameworks and their implementation; and the policy and legal challenges to national and local governments to help boost regulations' change for e-Governance success and support progress towards a better understanding of legal and regulatory issues and barriers in e-Governance.

7. CONCLUDING REMARKS

One of the clearest conclusions emerging from this study is that state structures, and institutional and legal factors, can be profoundly important in determining the nature, cost and success of e-Governance. Cohesive legislative and regulatory frameworks are necessary to bolster government effectiveness in improving governing, advancing democracy and fostering transparent

decision-making and inclusive polices in the local communities. The challenge is to promote legal changes and regulatory conditions, which provide for back office modernization, inclusive access, trust and confidence, and better public sector information to all. This, however, requires synchronized action in the administrative, data protection, privacy and security, and information provision laws involved. Coordination between the European, national and regional legislation is necessary to allow public administrations develop effective electronic services and development policies in the spirit of the main principles of Good Governance. The challenge, however, is to encourage governments to foster the regulations and the mechanisms ensuring the IT infrastructure, access and content that would best serve their citizens and the local communities. Much more importantly, there is sufficient evidence today that e-Governance development relies on the development of clear and comprehensive legal framework spanning across e-Government, e-Business and e-Democracy which helps government improve its performance continuously and build public confidence and participation. Progress clearly requires consideration of both organizational and technological conditions which often reflect national and/or regulatory and institutional regimes and cultures. Promoting legal changes hand in hand with service and technology innovation and government modernization is much essential to the success. Implementation of these laws and principles can assist the process of e-Governance in the European nations and the local communities. Though success in e-Governance can be often attributed to strong central leadership, consisting of an overall vision, strategies, roadmaps, resources and the specification of standards and frameworks, this needs to involve the local and regional initiatives, close to their social and business communities, who are able to strike a balance between the diverse interests and the local autonomy and freedom responsive of the specific local needs. Different localities may develop their own paths reflective of their unique identities, legal and institutional structures and cultures while, however, complying with national and international standards to ensure integration and interoperability. To succeed, national and regional legislative frameworks relevant to e-Governance must continue to improve and coordinate in the spirit and main principles of Good Governance to reach the objectives of the Knowledge and Networked society in the 21st century.

8. REFERENCES

- [1] European Commission. EGovernment Communication, http://europa.eu.int/information_society/eeurope/2005/all_about_egovernment/communication/text_en.htm, 2003.
- [2] Ibid.
- [3] Ibid.
- [4] Lisbon European Council. Presidency Conclusions, March 23-24, 2000. http://www.europarl.eu.int/summits/lis1_en.htm,
- [5] 5th Worldwide Forum on e-Democracy, Issy-les-Moulineaux, Paris, France, <http://www.issy.com/Rub.cfm?Esp=1&Rub=19>, September, 29-30 2004.
- [6] The World Bank, eGovernment Programme, <http://www1.worldbank.org/publicsector/egov/>.
- [7] European Union, Fourth European conference on e-Government, Dublin, Ireland, 17-18 June, 2004.

- [8] 5th Worldwide Forum on e-Democracy, Issy-les-Moulineaux, Paris, France, <http://www.issy.com/Rub.cfm?Esp=1&Rub=19>, September, 29-30 2004.
- [9] http://europa.eu.int/ISPO/basics/i_europe.html
- [10] http://europa.eu.int/information_society/programmes/egov_rd/documentation/text_en.htm#recommendations
- [11] Paskaleva-Shapira, K. 'INTELCITIES: eGovernance Platform and Research Approach', in Di Maria, E. (et. al) 'e-Governance practices, strategies and policies: State of the art', EU 6FP IP INTELCITIES Project Report 11.1.2: www.intelcitiesproject.com, 2004.
- [12] Paskaleva-Shapira, K. The regulatory conditions and legal framework as an eGovernance enabler, in Ásta Þorleifsdóttir (et al), 'City eGovernance: Best Practice Report', EU 6FP IP INTELCITIES Project Report 11.2.2, www.intelcitiesproject.com, 2004.
- [13] Lenk, K., Traunmüller, R., and M. Wimmer. The Significance of Law and Knowledge for Electronic Government, in: Grönlund, A. (ed.), Electronic Government – Design, Applications and Management, Hershey, Idea Group Publishing, 2002.
- [14] McNeil, R.S. (et al). Innovating in Digital government in the American states. Social Science Quarterly, vol. 84 No. 1, March 2003.
- [15] OECD. Checklist for eGovernment leaders, http://www.oecd.org/findDocument/0,2350,en_2649_34131_1_1_1_1,00.html, 2004.
- [16] Amoussou-Guenou, R. Legal aspects of eGovernment, United Nations' Regional Workshop on eGovernment, Bangkok-Thailand, May 31-June 2, 2003.
- [17] Millard, J. and S. Iversen (Ed.). Reorganization of government back-offices for better electronic public services – European good practices, Final report to the European Commission, p. 61: http://www.oeaw.ac.at/ita/ebene5/back_office_reorganisation_volume1_mainreport.pdf, 2004
- [18] European Commission. Report on the Implementation of the EU Electronic Communications Regulatory Package, http://europa.eu.int/information_society/topics/telecoms/regulatory/new_rf/index_en.htm#Introduction, 2003.
- [19] Government definitions on the web: <http://www.google.com/search?hl=en&lr=&oi=defmore&q=define:government>.
- [20] Official Journal L 281, 23 November 1995, p. 31.
- [21] Official Journal L 24, 30 January 1998, p. 1.
- [22] Official Journal L 13, 19 January 2000, p. 12.
- [23] Official Journal L 178, 17 July 2000, p. 1.
- [24] Official Journal L 201, 31 July 2002, p. 37.
- [25] European Commission, The Amsterdam Treaty: a Comprehensive Guide, <http://europa.eu.int/scadplus/leg/en/s50000.htm>.
- [26] Millard, J. and S. Iversen (Ed.). op.cit.
- [28] <http://www.hrweb.org/legal/cpr.html>
- [29] <http://conventions.coe.int/Treaty/EN/cadreprincipal.htm>.
- [30] European Commission. Report on the Implementation of the EU Electronic Communications Regulatory Package. Op. cit.
- [31] Accenture. EGovernment Leadership-Realizing the Vision Report, http://managementconsult.profpages.nl/man_bib/rap/accenture07.pdf, 2003.
- [32] Reichstadter, P. E-Delivery – based on the Austrian e-Government Law', e-Challenges 2004 Conference, Vienna November 7-8, 2004.
- [33] Accenture. 2003. Op. cit.
- [34] Millard, J. and S. Iversen (Ed.). Op.cit.
- [35] Accenture. 2003. OP.cit.
- [36] Ibid.
- [37] Brown, D. and M. Isakovic. Security and privacy laws: Striking the balance'. Report on a seminar on Commonwealth Electronic Governance: http://www.electronicgov.net/pubs/workshop_reports/security-privacy03.shtml, 2003.
- [38] Results of analysis of the eGovernment fact sheets by countries: Legal framework, IDA eGovernment Observatory, <http://europa.eu.int/ida/en/chapter/424>.
- [39] UK Office of deputy Prime Minister, Local e-Government, UK eGovernment News – 20 May 2003.
- [40] de Montigny, Y., in: UK Office of the e-Envoy (OeD). 'Report on eGovernment', p. : <http://www.e-envoy.gov.uk/assetRoot/04/00/08/23/04000823.pdf>, 2004.
- [41] Brown, D. and M. Isakovic. Op.cit.
- [42] http://www.csis-scrs.gc.ca/eng/backrnd/back12_e.html.
- [43] <http://www.info.gov.hk/english/itleto.htm>.
- [44] UK Office of the e-Envoy (OeD). Report on eGovernment, p.213: <http://www.e-envoy.gov.uk/assetRoot/04/00/08/23/04000823.pdf>, 2004.
- [45] <http://www.qlinks.net/comdocs/elsig/>
- [46] UK Office of the e-Envoy (OeD). Op. cit.
- [47] Leitner, C. et al (Ed.). eGovernment in Europe: The State of Affairs. Maastricht: European Institute of Public Administration, 2003.
- [48] Di Maria, E. (et. al) 'e-Governance practices, strategies and policies: State of the art', EU FP6 IP INTELCITIES Project Report 11.1.2: www.intelcitiesproject.com, 2004.
- [49] UK Office of the e-Envoy (OeD) Op. cit.
- [50] Brown, D. and M. Isakovic. Op.cit.
- [51] Ibid.
- [52] European Union, Information Society. eEurope 2005 Security Policies in Brief, 2004.
- [53] Brown, D. and M. Isakovic. Op.cit.
- [54] Ibid.
- [55] European Union, eGovernment Good Practice Framework, 2004.
- [56] European Union. The Role of eGovernment for Europe's Future, 2003.

A National Center for Digital Government

Program on Networked Governance

Project Highlights, dg.o 2006

NSF Grant # 0131923

Jane E. Fountain

University of Massachusetts Amherst,
National Center for Digital Government
[jfountain@polsci.umass.edu](mailto:jfountain@polisci.umass.edu)

David Lazer

Kennedy School of Government, Harvard University
Program on Networked Governance
david_lazer@harvard.edu

ABSTRACT

In this paper, we describe the ongoing research and activities of the National Center for Digital Government, now based at the University of Massachusetts Amherst, and the Program on Networked Governance, the successor program to the NCDG, based at the Kennedy School of Government at Harvard University.

General Terms

Digital government, e-government, public management, social network analysis, political analysis, institutional analysis

Keywords

Digital government, e-government, public management, social network analysis, political analysis, institutional analysis

1. ACCOMPLISHMENTS: RESEARCH

The central projects in progress this year include the DNApolicy.net project, an examination of web-based information gathering by state and local government officials; Neteam, a study of the role of ICT in communication on teams; the Data Mining Project, an evaluation of the potential to mine DNA databases for relatives; a study of state health officials and the role of ICT in knowledge diffusion; the production of an edited volume on information government, Igov; continued data collection for the Connecting to Congress Project; continued data collection for the Project on sustainable interorganizational relationships in government; the first phase of data collection for the U.S. Japan comparative study of political development in central governments; and replication and use of survey and interview protocols for the study of trade administration modernization using ICT in the Government of Mexico.

Fountain has synthesized the results of the interorganizational relationships in government study into the multi-level integrated information system (MIIS) framework [1]. The framework draws from economic sociology and new institutionalism to develop a model of organizational change in government that encompasses individual and small group interaction at the team level and network level, organizational and interorganizational routines and procedures, and institutional norms and overarching frameworks that constrain and influence behavior. In related work, Fountain reported the finding of two cross-agency initiatives, Grants.gov and e-Rulemaking [2]. Lazer has produced papers examining the impact of efficient information diffusion on systemic performance, as well as on the potential for data mining DNA databases.

2. ACCOMPLISHMENTS: OUTREACH

2.1 Colloquia, Invited Lectures and Keynote Addresses

During 2005-06 Lazer supervised a colloquium series with more than ten invited speakers and initiated the Trans-Atlantic Interactive Conference (TAICON) to bring together researchers in the Boston and Cambridge communities with those at ETH Zurich. Lazer started a blog on complexity and social networks and supervised 6 research fellows. Fountain is hosting four fellows, including two faculty fellows from the Netherlands and Korea. She webcast an invited keynote address to the FP7 eGovernment Research Stakeholder Consultation Workshop, in Brussels; traveled to Wellington, New Zealand, to deliver a keynote address at the Knowledge Management Asia Pacific 2005 Conference which was co-hosted by the School of Information Science and the School of Government at Victoria University, Wellington. She was the only U.S. invited lecturer at the NetGov meeting, an EC eGovernment planning session held at the Gabriel Lippmann Research Center in Luxembourg.

2.2 Collaboration: Success Stories

Among the success stories for the 2005-06 year are the following: David Lazer, director of the Program on Networked Governance, collaborated with the Ash Institute at the Kennedy School to reach 150 people at more than 30 labs through a video conference. In addition, the Swiss Consulate continues to support the joint PNG-ETH Zurich online colloquia (TAICON) which bring together U.S. and European researchers.

Fountain had shared research instruments and protocols with researchers in Mexico, Peru, Canada and Japan in order to disseminate methodologies for institutional analysis of digital government developments. The U.S.-Japan comparative research study explicitly builds on the technology enactment model as well as knowledge sharing between the University of Tokyo, 21st Century Center for Excellence on Policy Systems in Advanced Countries and the National Center for Digital Government at the University of Massachusetts.

2.3 Collaboration: Partners

Within Harvard University, the central partners for collaboration this year have been the Institute for Quantitative Social Science, the Center for Business and Government, the Taubman Center for State and Local Government, and the Ash Institute for Democratic Governance. At the University of Massachusetts Amherst, central collaboration partners include social science and computer science research entities: the Center for Public Policy, the College of Social and Behavior Sciences, the Center for Technology and Dispute Resolution (based in the Legal Studies Department), the IBM Linux-Open Source Software Teaching Lab (based in CSBS), and the Electronic Enterprise Institute (based in the College of Natural Sciences and Mathematics and the School of Management).

Our partnerships with other universities continue to catalyze new research and outreach activities. These partners include: the University of California, Ohio State University, MIT (Media Lab and CSAIL), ETH-Zurich, the University of Amsterdam, the University of Tokyo, Victoria University in Wellington, New Zealand, the University of Utrecht, and the Seoul Development Institute in Korea.

Research with government entities continues with partnerships involving DNA laboratories, the U.S. House Committee on Administration, the Presidential Management Agenda 25 cross-agency egovernment projects; and the Office of IT and Egovernment in the U.S. Office of Management and Budget. Finally, we partner with the Congressional Management Foundation, a nonprofit research and training institution, through the Connecting to Congress Project.

3. BROAD IMPACT

NCDG and PNG constitute physical, institutional and informational resources for research, education, community building and outreach in several countries. They are explicitly community building and network building activities which strengthen and sustain an emergent research base in science and technology for digital government research. NCDG and PNG Fellows are placed in universities throughout the world. Web-based content, papers, seminar presentations, workshop reports, webcasts, and video presentations made accessible through the NCDG and PNG websites provide resources for the DG community and researchers throughout the globe. Finally, the social science studies of digital government, which connect to mainstream, disciplinary research, are deepening the DG research program and helping to embed DG research in university departments and curricula.

4. CONCLUSION

The NCDG and PNG continue energetic and productive research, outreach and infrastructure building programs that are educating and connecting research fellows, forming international comparative research projects and knowledge sharing opportunities through colloquia, seminars, and web-based activities, and continuing to produce academic, published research on the development, outcomes, and implications of digital government.

5. REFERENCES

- [1] J. E. Fountain, "Challenges to Organizational Change: Multi-Level Integrated Information Structures (MIIS)," in D. Lazer and V. Mayer-Schoenberger, eds., *Information Government* (under review with MIT Press, forthcoming).
- [2] J. E. Fountain, "Central Issues in the Political Development of the Virtual State," in M. Castells and G. Cardoso, *The Network Society: From Knowledge to Policy* (Brookings Institution Press, 2006 in press).

Connecting to Congress

David Lazer

Program on Networked Governance

Harvard University

david_lazer@harvard.edu

Kevin Esterling

Department of Political Science

University of California

kevinesterling@ucr.edu

Michael Neblo

Department of Political Science

Ohio State University

neblo.1@osu.edu

Jane Fountain

National Center for Digital Government

Program on Networked Governance

University of Massachusetts

Fountain@polisci.umass.edu

Ines Mergel

Harvard University

ines_mergel@harvard.edu

Curt Ziniel

Department of Political Science

University of California

curt.ziniel@ucr.edu

ABSTRACT

In this paper we summarize the progress of the Connecting to Congress (CTC) project.

General Terms

Management, Performance, Human Factors, Theory.

Keywords

Information Technology, Congress, Diffusion.

1. RESEARCH ASPIRATIONS

The Internet has the potential to transform our democracy—a potential that has begun to receive substantial scholarly attention. This attention has focused on the potential transformational effects of the technology on civil society, and, in the political realm, how the Internet might transform political discourse (e.g., DiMaggio & Powell, 1983). Researchers have devoted little attention, however, to how the Internet might transform existing institutions for connecting citizens to elected officials. This relationship is the fundamental building block of a representative democracy, and it has come under increasing strain as our country has grown from a few million to a few hundred million; as congressional districts have swelled from a few tens of thousands to well over six hundred thousand; as the number of matters the state is involved in has multiplied; and as policy problems have grown more complex. Contemporary Washington politics is now almost exclusively the domain of entrepreneurial legislators, highly trained committee staff, legal counsel, agency heads, lobbyists, and expert policy analysts. Today, it is difficult for interested citizens to even understand the policy process, much less have their voice heard in it (Hecko, 1974; Lupia & McCubbins, 1998). As a consequence of this and other trends, citizens have become increasingly disengaged from the work of Congress. The Internet offers a set of tools that might help to arrest this trend, and to fundamentally alter the level of participation of citizens in the consultative process with their Representatives. A well-designed Internet strategy by Members of Congress can provide citizens with information useful for understanding a policy as it develops, while also allowing citizens to interact more symmetrically with both their Member of Congress and with each other. Wisely used, the Internet can re-connect citizens and Congress.

Further, the House of Representatives offers a unique setting for understanding the opportunities and obstacles to effective use of information technologies in the public sector. Congressional offices function as 440 small, functionally identical, and independent public organizations (Salisbury & Shepsle, 1981). This decentralization enables a large N statistical study of innovation adoption, in essence to test our expectations that the behaviors of 7 Members of Congress can be explained by their recent electoral experiences, district characteristics, institutional resources (e.g., Fenno, 1978), and embeddedness in social networks (Walker, 1973).

2. COLLABORATIONS

The foundation of the CTC project is a collaboration among researchers at four universities (Harvard, UMass, Ohio State, and University of California) and the Congressional Management Foundation (CMF). The CMF is a small nonprofit organization devoted to improving the management of the institution of Congress. With the support of the Pew Foundation, they issued a series of influential reports on the use of the Internet by Members of Congress. As an organization, they have remarkable insight and depth of connections into the institution. The team of researchers involves four faculty and a postdoc with complementary knowledge and skills: Social network analysis (Lazer, Mergel); Institutional analysis (Fountain); Quantitative methodologies (Esterling and Neblo); Congress (Esterling); and Deliberative processes (Neblo).

3. ACCOMPLISHMENTS 2005-2006

Our primary accomplishments are around the collection of data to address the research questions enumerated above. We would highlight three data gathering efforts.

3.1 Collection of quantitative data

We conducted a comprehensive coding of all Member websites in 2005, along the dimensions of issue content, technological sophistication, and constituent services. Further, we have gathered a fairly large array of complementary quantitative information on Members and their districts. Our analyses (from which we have one published paper and one forthcoming) indicate that tenure and electoral security strong predictors of the quality of members' websites. We are also working on linking these data to a variety of variables for each Congress that measure differences in the information flows that Members' offices are exposed to, including: overlap of

committee membership; adjacency of districts; and proximity of offices. Initial analyses suggest that surprisingly little diffusion of web practices occur through informal networks within Congress.

We are also beginning to take a subset of these codes to develop longitudinal data on features of websites, using the Internet Archive.

To support some of the data gathering effort, we have also developed the “AskThomas” data gathering tool, which enables the rapid extraction of data from Congress’ Thomas search engine. Thomas is designed to produce data on specific queries about particular Members of pieces of legislation. The AskThomas tool will, for example, produce spreadsheets arranging all legislators against all pieces of legislation, with codes for whether a legislator co-sponsored and/or voted for a particular piece of legislation. We anticipate that this tool will be widely used by scholars.

3.2 Qualitative research

We have also begun a large scale effort at interviewing Congressional staff, focusing on the Communications Directors within Congressional offices, as well as conducting in depth case studies in a small number of offices. The objective of these analyses will be to understand the internal processes associated with effective use of the Internet by Members. Initial results from the qualitative research highlights the role of a technological “entrepreneur” within an office, as well as the critical importance of a number of small vendors that have emerged to serve Congressional offices. The interviews also suggest that while Members from marginal districts may be more aggressive in using the Internet, they are very cautious about the type of policy information they place on their sites, because of the constant possibility of alienating constituents. This suggests a hypothesis that we are planning to test with the development of new codes for Member sites: that Members from competitive districts have more content on their sites, but less clarity.

3.3 Experimental research

We are also in the process of designing the protocols for a series of experiments we will be conducting in the Spring. These will be “deliberative experiments” in which Members will interact online with a small group of constituents. Our objective will be to vary the information presented to constituents as well as the nature of the forum in order to examine what affects key deliberative outcomes (e.g., opinion change, tolerance of opposing points of view, etc).

4. MANAGEMENT STRUCTURE

The project is coordinated through the *Program on Networked Governance*. We have a project website on which we place all key documents. We also have a monthly conference call reviewing progress on each component of the project, and the key project leaders meet at least 3 times a year at conferences and DC.

5. COLLABORATION SUCCESSES

This project leverages a unique collaboration anchored at the PNG at Harvard University, among academic political scientists, information technology researchers, CMF supplemented with the cooperation of the U.S. Congress itself. This collaboration combines deep substantive knowledge,

research training, organizational capacity, and a ready access to Congress into a synergistic relationship, one that offers a research potential that is much greater than the sum of its parts.

6. BROAD IMPACT

The potential of this project is to better understand how the Internet might be useful to Members of Congress, and, in so doing, affect for the better how they use the Internet.

7. OBSTACLES/CHALLENGES

The major challenge to the project is the lack of institutional memory within Congress. While we can use the Internet Archive to trace the changes in the usage of the Internet by Members over time, the velocity of events and turnover among staff means that what happened even a year ago is ancient history. Thus, attempting to understand the choices by offices in 1999 has been difficult.

8. RESEARCH VALUE OF DG DOMAIN

We would highlight two key values: (1) that the use of technology by individuals and organizations offers a prism through which to view individual and organizational behavior, and (2) that digital technology, in particular, often leaves traces of human behavior (e.g. websites) that might otherwise disappear.

9. ADVICE TO DG PROGRAM

The fundamental challenge to the DG program is how to enable collaboration across disciplinary boundaries. One possible way to do this might be to have some type of “briefings” by social and computer scientists providing a “lay of the land” of their respective corners of academia.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support of NSF grant No. 0429452. Updates on the progress of the research will be reported at www.ksg.harvard.edu/netgov.

Digital Deliberation: Searching and Deciding About How to Vote

Scott P. Robertson

Drexel University

College of Information Science and Technology

Philadelphia, PA 19147-2875 USA

+1 215-895-2476

scott.robertson@drexel.edu

ABSTRACT

This paper summarizes a new NSF-funded research project to study and develop an online portal that supports voter deliberation and decision making. User-centered design methods with varied population groups will be employed to develop features and test prototypes of a voter portal.

Categories and Subject Descriptors

H.5.2 [Information Interfaces and Presentation]: User Interfaces—User-centered design; prototyping; H.5.3 [Information Interfaces and Presentation]: Group and organization interfaces—Web-based interaction. K.4.1 [Computers and Society]: Public policy issues

General Terms

Design, Experimentation, Human Factors

Keywords

Digital government, electronic voting

1. INTRODUCTION

A considerable amount of political activity is moving onto the Internet. This includes information dissemination by candidates, parties, issue advocates, campaigns, governments, news and opinion media, and individuals, all of which is directed at citizens and potential voters. It also includes active participation by citizens in interactive digital environments mediated by social software such as blogs, opinion forums, discussion groups, etc. While there has been considerable study of the large scale demographics of this movement primarily using survey methods, there have been few “close up” studies of individuals involved in digital democracy using experimental and observational methods. There has also been a lack of empirical user studies in support of the design of software interfaces to support digital democracy.

The screenshot shows a web-based voter portal. At the top right, there are dropdown menus for 'Focus on' (set to 'All Categories'), 'and' (set to 'All Ballot Items'), and 'and' (set to 'All Candidates'). Below this is a search bar with the placeholder 'Notes: Your 2' and buttons for 'VIEW' and 'NEW'. The main area displays a table of ballot items. The columns are: Ballot Item, Candidate/Ballot, Title, Source, Category, and Date. The table contains several rows, including:

Ballot Item	Candidate/Ballot	Title	Source	Category	Date
Proposition 49 Before and After School Programs	Opposes California State Proposition 49 Before and After School Programs	Minion Hunter - Berkeley City Council	Other	24/09/2002	
U.S. Senate Medea Benjamin	Medea Benjamin on the Issues	OnTheIssues.org	Other	03/06/2000	
Proposition 49 Before and After School Programs	Summary	CA Secretary of State	Official	18/08/2002	
U.S. Senate Medea Benjamin	Candidate Statement	CA Secretary of State	Official	18/08/2002	
Secretary of State Keith O’Leary	Candidate Statement	CA Secretary of State	Official	18/08/2002	
Proposition 49 Before and After School Programs	Arguments	CA Secretary of State	Official	18/08/2002	
Proposition 49 Before and After School Programs	Tell	CA Secretary of State	Official	18/08/2002	
U.S. Senate Medea Benjamin	Open Seats Candidate Votes the Party Line	The Washington Observer	News	14/03/2003	

Below the table, there is a section titled 'Proposed Proposition Forty Nine (49) Before and After School Programs: State Grants Initiative Statute' with a logo for the 'Official Voter Information Guide'. It includes sections for 'ARGUMENT in FAVOR of Proposition 49' and 'ARGUMENT Against Proposition 49', each with bullet points. A note at the bottom states: 'Arguments printed on this page are the opinions of the authors and have not been checked for accuracy by any official agency.'

Figure 1. A Voter Portal would filter and sort political information from the Internet and allow users to browse, compare, annotate, and share opinions about items.

In this project, people will be observed closely as they use digital materials to make voting decisions. A series of empirical studies using real online materials in mock voting exercises will examine user behavior in this domain. Information gathered in the empirical studies will be used to help design a Voter Information Portal (Robertson, 2005) that organizes political information for citizens and helps them make decisions. The portal will use Internet information found by commonly-used search engines and allow voters to reorganize it into user-centered categories, annotate it, and share it. Issues of privacy and ethics in this context will be examined.

2. PRELIMINARY STUDY

An experiment was recently completed which tested the use of a portal. Participants completed a mock voting exercise in which they studied real campaign materials for historical elections with which they were unfamiliar (a California Senate race, a California Secretary of State race, and a California ballot proposition) and then voted electronically. *Information Presentation* was varied, with subjects receiving the materials on paper in one condition and via an electronic portal in another condition. In the Paper condition, materials were organized by categories into notebooks with indexes and tabs. In the Electronic condition, materials were organized by the same categories visible in navigation hyperlinks. Crossed with the *Information Presentation* factor was a *Ballot Integration* factor. In Ballot-Integrated conditions the electronic

ballot was available during the information searching phase whereas in the Ballot-Not-Integrated condition the ballot was only available after information search

To summarize the results:

- The notebooks never had any advantage over the electronic portal, and the portal seemed to help people use the materials more easily and quickly. The ease of use of the portal may have contributed to a better understanding of the materials.
- The availability of the ballot during information browsing never had any advantage, and in fact subjects found it easier to complete their tasks of studying materials and deciding on their choices when voting took place after the study phase.
- The enhancing effect of separating browsing and voting was greatest, as measured by “ease of making a final decision” ratings, when the electronic portal was being used.
- There are strong order effects in information browsing and thus the design of an information portal will influence how voters study and learn about issues.

These results are consistent with a view that voters wish to separate study, deliberation, and choice making from voting, perhaps because voters wish to maintain a minimal cognitive load by focusing on one thing at a time. Once voters begin the process of casting the ballot, they prefer to have their choices already made. If they browse and vote at the same time, they will “chunk” their activity such that all deliberation about a particular issue or candidate is finished and the vote cast before moving on. The need to manage cognitive and memory load factors may be even greater when a portal is being used to the degree that the portal itself contributes to cognitive load either because of its design or the amount of information that it can present in a short time.

3. PROPOSED STUDIES

In a set of studies over the next three years, the investigators will examine the conceptual categories that people use to understand political information, browsing and decision making strategies, and personalization. Another set of studies will examine more social uses of digital information. Questions include interest in sharing profile information, annotations, and opinions with others, impact of shared information on decision making, and the interaction of privacy and information sharing. All studies will compare participants from diverse backgrounds and situations. In each case there is a preliminary study based on a participatory design method (Schuler & Namioka, 1993) such as the focus group which is then followed by an empirical, mock voting study.

A practical outcome of these studies will be design concepts for a political information browsing portal that supports all of these activities. A series of user-centered prototyping sessions (Alvarez & Hall, 2004) will be carried out with the goal of producing a usable political browser prototype by the end of the second year. The last year of the proposed project period (2008) is an election year. A prototype system will be in place early in the year and made available to a large number of voters of varying backgrounds via the Internet. Portal users will be volunteers who agree to have their usage behavior monitored (anonymously).

After the election, a survey of users will be conducted to assess usability, interest in using a real working system in future elections, and concerns.

4. CONCLUSION

Putnam (2000) argues that U.S. culture is undergoing a large-scale loss of “social capital,” or interpersonal networks of trust and knowledge. This project may infuse some components of social capital into the electoral decision making process by supporting browsing, collaboration, and knowledge sharing. It may help voters to organize their information seeking behaviors and integrate them during the decision process leading up to elections.

Researchers and developers should work against the development of an electronic voter support system that leverages the “digital divide” to further exclude people who are not well integrated into the socioeconomic fabric, and that further exacerbates the problem of alienation by not including community-building components. By taking a contextual, user-centered perspective we have an opportunity to affect the democratic process in the future. The design of electronic voter support systems must be informed by an HCI-perspective which includes an understanding of users, broadly defined user behavior (including affective, cognitive, and collaborative behaviors supporting information gathering and decision making), and context.

Democratic government, and the technologies that will support it in the future, need to be led by the needs and requirements of citizens, not by the capabilities and features of emerging technologies. There is considerable research in political science, psychology, and sociology on political decision making. There is less research on how these learning and deliberation processes are transferring as new digital information sources and virtual deliberation tools become available to citizens. The goal of this research is to shape new technologies for democracy as they emerge.

5. ACKNOWLEDGEMENTS

This work is supported by the National Science Foundation under grant number IIS-0535036. Thanks to Palakorn Achananuparp, Jim Goldman, Sang Joon Park, and Nan Zhou for their work on this project to date.

6. REFERENCES

- [1] Alvarez, R.M. & Hall, T. 2004. *Point, click, and vote: The future of Internet voting*. Washington, DC: Brookings Institution Press.
- [2] Putnam, R. 2000. *Bowling alone: The collapse and revival of American community*. New York: Simon & Schuster.
- [3] Robertson, S. 2005. Voter-centered design: Toward a voter-centered decision support system. *ACM Transactions on Computer-Human Interaction*. 12(2), 263-292.
- [4] Schuler, D. and Namioka, A. (Eds.). 1993. *Participatory design: Principles and practices*. Hillsdale, NJ: Lawrence Erlbaum Associates.

SESSION 6A

STUDENT RESEARCH PRESENTATIONS

Moderator

Sharon S. Dawes, University at Albany/SUNY, USA

Titles and Authors

The Social Relations of e-Government Diffusion in Developing Countries: The Case of Rwanda
Mwangi, Wagaki

*e-Governance in Africa, from theory to action: a practical-oriented research and case studies on
ICTs for Local Governance*
Misuraca, Gianluca Carlo

Automated Classification of Congressional Legislation
Purpura, Stephen; Hillard, Dustin

The Social Relations of e-Government Diffusion in Developing Countries: The Case of Rwanda

Wagaki Mwangi

Syracuse University

Department of Political Science, Syracuse, NY 13244-1020

Tel +1-315-443 2416

ermwangi@maxwell.syr.edu

ABSTRACT

Rwanda has undergone a rapid turnaround from one of the most technologically deficient countries only a decade ago to a country where legislative business is conducted online and wireless access to the Internet is available anywhere in the country. This is puzzling when viewed against the limited progress made in other comparable developing countries, especially those located in the same region, sub-Saharan Africa, where the structural and institutional constraints to e-government diffusion are similar. Based on an exploratory case study of the country's e-government system that draws on group and social theories, I argue that the convergence of four factors associated with the policy environment, political leadership, emigrants and refugee returnees, and epistemic communities account for Rwanda's achievements. The primacy of interest group politics in the unfolding story of e-government diffusion in developing countries is underscored and potential areas for further research highlighted.

Keywords

E-government diffusion, norms, social theory, Rwanda, sub-Saharan Africa.

1 INTRODUCTION

Only a decade after emerging from the fastest genocide of the 20th Century, Rwanda, a small country in Eastern Central Africa, has become one of the continent's leaders in, and model on, bridging the digital divide through e-government. By August 2005, Rwanda was the only African country where official business by members of the legislature – the House and Senate – was conducted online, including voting on motions and the circulation of bills and documents [11]. By February 2005, every part of the country had access to the Internet through radio-wave (wireless) connection, and the deployment of fiber optic across the country was well under way. Nearly all public agencies were networked. In June 2005, Rwanda completed the review of implementation of its first five-year national information and communication infrastructure (NICI) plan, and embarked on preparing the second plan. In contrast, 25 African countries were

still in the process of preparing their first plan, among them Kenya, Uganda, Zambia and Zimbabwe; countries which were already connected to the Internet in 1995 when Rwanda was still reeling from the effects of genocide [1]. Moreover, diffusion in countries such as Zambia that were among the first to acquire full Internet access in 1993 have nearly stagnated. How do we make sense of Rwanda's leap forward?

This study has value for both policy and theory. The international community is unrelenting in its commitment to bridge the digital divide. To this end, the eighth goal of the United Nations Millennium Development Goals adopted in 2000 calls for cooperation with the private sector in order to make available [by 2015 to those two million without] the benefits of new technologies – especially information and communications technologies. Accordingly, several initiatives are in progress in the developing, and least developed countries, despite a limited understanding as to how various factors interact to produce progress in some countries and retreat or stagnation in others. This study presents a framework that can expose the social and political dynamics that mediate in e-government diffusion efforts.

The study also opens up new avenues for research by rendering a political understanding to an issue that is primarily viewed as scientific-technological and administrative. When e-government diffusion is viewed as a socially constructed political problem, we are able to better understand the messiness of e-government diffusion in developing countries, and to understand why structural and institutional factors offer insufficient explanations.

Through an exploratory case study of e-government diffusion in a country that in under a decade has moved from a technology laggard to leader renders explicit the role of agents in making the turnaround possible. The study diverges from previous approaches that predominantly focus on diffusion failures in developing countries. Rather, I examine a case that has made headway despite difficult conditions, and assess the extent to which the elements identified interrelate and are applicable elsewhere especially in sub-Saharan Africa. The study is the product of archival research conducted between Fall 2004 to Fall 2005, and a month-long residence research in Rwanda in June and July 2005 that involved 21 in-depth interviews with politicians and key informants from government, business, academia, and voluntary not-for-profit organizations.

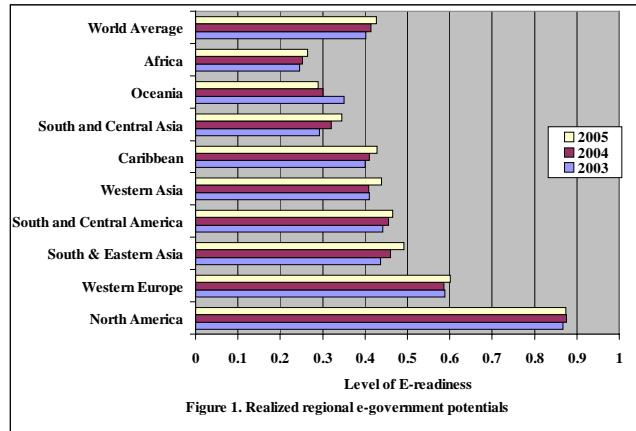
This paper proceeds as follows. I present trends in e-government diffusion and review key literature on e-government in particular as it pertains to Africa, and then highlight my research design. This is followed by a presentation of the findings and then a

discussion of the their implications for practice and theory, and ideas for further research. Details of my research methods are contained in an appendix at the end of the paper.

In this paper, I use the terms Internet, information technology and information and communications technology interchangeably. The concept refers to the entire range of technologies that may be used to set up an information system to facilitate all forms of data exchange in public agencies through the Internet. So how does all this help to make sense of Rwanda's e-government diffusion leap forward?

2 DIFFUSION TRENDS

E-government refers to the provision of services to citizens by public agencies through the use of Information Technology (IT). Initiatives to explain e-government trends across countries are diverse and difficult to compare [2] but country rankings that measure e-readiness although new, offer a comparable basis to view developments across regions and countries. Although variation differs across countries as demonstrated by the 2005 United Nations Report [14], regional trends over time are fairly consistent (see Figure 1 below).



At the turn of the century, e-readiness rankings started appearing [2]. Of these, the UN rankings are perhaps the most useful measures of progress to date because determining what exactly is measured, and how, is political and negotiated between governments in advance. Therefore, the results above constitute the consensus by governments of what constitutes e-government, and how its diffusion has varied across regions during the last three years.

The UN's e-government ranking is based on a country's e-readiness (capacity dimension) and e-participation (willingness dimension) and captures the potential for e-government that is already realized. This study focuses on the capacity factors, the diffusion dimension, not adoption (willingness). The UN's three indicators of e-readiness are web measure, telecommunications infrastructure and human capital. The web measure captures the sophistication of a country's online presence. Human Capital is a composite measures of education derived from the combination weighted ratio of adult literacy and academic enrollment from elementary to tertiary levels. The index on telecommunications infrastructure includes a variety of indicators such as access to personal computers, telephone lines, and online population.

The UN report hypothesizes that there "appears" to be a strong relationship between e-government readiness and income *per capita*, education, and infrastructure development, and that part of the high e-readiness in developed countries is due to their past investment in, and development of, infrastructure. There are two significant things about this proposition. First, it lacks specificity about the causal nature of this relationship, that is, whether it is correlational, covariate, exponential or some other kind. Second, it is consistent with the conclusions reached about IT diffusion across countries in the 1990s, where regional disparities in the *per capita* IT user base were also found to be due to prevailing structural conditions such as poverty, illiteracy, human resource scarcity, and poor infrastructure, including telephony, electricity, and technological deficiency, as well as political factors [7-9, 12, 13]. Still, income remained the best predictor of IT diffusion [15]. However, both the e-government and IT diffusion findings are problematic.

While broad generalizations seemed to make sense across countries in which there are vast disparities in these structural variables, their effects were less certain when countries were examined after controlling for regional disparities. For example, empirical evidence of e-government diffusion in sub-Saharan Africa, where structural and institutional conditions are fairly similar, reveals no consistent patterns associated with income *per capita*.

Figure 2 below shows that an individual's income level is not a strong determinant of e-readiness among the 48 sub-Saharan African countries examined. No doubt economic indicators matter, but they do not tell the whole story.

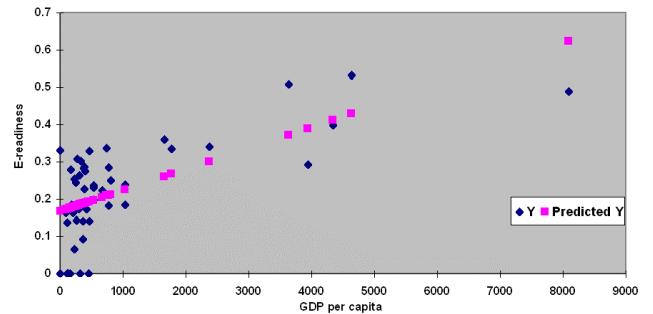


Figure 2: Relationship between income and e-readiness in sub-Saharan Africa

Critics argue that e-readiness rankings postulate a 'one-size-fits-all' e-government and disregard the characteristics, investment, and application requirements of individual countries [2]; that is, they ignore the context. This is echoed by studies that suggest that e-government in Africa has failed due to a superimposition of incompatible western models on developing country conditions [6]. They emphasize the need for a targeted e-government approach instead [17].

Moreover, the findings of the UN study appear rather inconsistent with what is on the ground. Rwanda's score on the infrastructure measure for e-readiness was estimated at 0.0035 in 2005 and 0.0040 in 2004, and was the defining variable in the country's low e-readiness ranking given its much higher ratings on the human capacity and web measure indices. Yet, by the time the UN study was conducted, 34 percent of all government and other public agencies had dial up connectivity, 38 percent were on leased line, 3 percent on VSAT and 41 percent on wireless. Further, 75

percent of all government agencies had websites, and over half (56 percent) had the Rwanda (.gov.rw) domain name. In terms of networking – the highest level of a five-rank web measure – 85 percent of the agencies were connected to the fiber optic backbone, and 41 percent had a wide area network, whereas 88 percent had a local area network [19]. How then do we explain this divergence in data that was collected within two months of each other?

Part of the answer lies in the indicators used to measure e-readiness. The UN report argues that the high level of e-readiness in developed countries is due to these countries' investment in infrastructure, especially technological and telecommunication infrastructure. The telecommunications infrastructure index used in the UN study is the weighted average of a country's infrastructure capacity, which is comprised of the number of people, per 1000, with personal computers, telephone lines, mobile phones, and televisions, as well as the online population and Internet users for every 1000 people. These measures too are problematic for two reasons. First, they focus on a country's *online* population – individual actors in the country – and not on the *institutional* population and their facilities, the appropriate measure for the bureaucrats offering the service. Moreover, it is unclear how an indicator such as television access is relevant to e-readiness. Second, the basis of these measures is telephony, yet Rwanda's IT system is based on a different backbone, radio waves, which is not captured by the UN indicators.

3 E-GOV: AS NORM AND POLICY

Thus, part of the problem in explaining e-government diffusion in developing countries is that the predominant theoretical and methodological approaches employed are problematic. But they also demonstrate a failure to take contextual factors seriously.

Be that as it may, I suggest that there is an even more important aspect of e-government that is recognized but glossed over by scholars of e-government – agency – which early sociology scholars of technology diffusion emphasized in their hypotheses of imitation [10], contagion and culture [18]. It is this emphasis on agency that my study builds on.

Using social theory reveals how the self-understandings and interpretation by Rwandese of their structural and institutional constraints converged and enabled them to establish a countrywide e-government system ahead of countries such as Senegal, Ghana, Kenya and Zambia, where the online and institutional populations were relatively high by the mid-1990s. It depicts how social and political influences constrain or enhance the alteration of the underlying structural and institutional structures, thereby interfering with the quantitative measures, and causal explanations sought.

Drawing on this perspective, I view e-government as a structural arrangement that is a spill over from a private sector practice known as e-commerce, whose aim was to enhance customer service effectiveness and efficiency. The norm emerged in developed countries first and its practice has not only assured its continuity, but also led to a spillover into the public sphere, as well as individual lifestyles, ranging from blogging, to online dating to personal communication.

I also follow Milner's hypothesis that as with any other policy, information technology, and by extension e-government, attracts interest group politics. Milner argues that political factors exert a strong influence on IT diffusion in that potential losers from the Internet will attempt to use political institutions to enact policies that block its spread [13]. I depart from her focus on political institutions *per se*, and pursue instead, similar arguments advanced by public choice theorists concerning interest group politics, agenda setting, and ideas. I assume that in developing countries where the e-commerce culture could not take root due to a lack of the requisite preconditions, e-government diffusion initially took the form of a policy goal that sparked bureaucratic and interest group politics before it was adopted as a norm. In these countries, the diffusion of e-government as a new norm would occur incrementally.

The implication of my study then, is not to refute the UN and others' claims that structural and institutional factors are not constraints to the establishment of e-government. Rather, it is to suggest that the meaning actors give to these constraints also matter in regard to whether or not they are able to establish an e-government system.

4 RELATIONS OF DIFFUSION

The first e-government efforts in Rwanda date back to 1983 when a government master plan for an information system was established, followed by the establishment of a national policy in 1992 [5]. But this mission was abandoned when the civil war broke out in 1993/4.

The effects of the war on the country's human resource and economy [5] were still evident by the end of 1997 when the latest e-government efforts began, three years after the new government took over in a *coup d'état*. At the time, 94 per cent of the country's population of 7.9 million lived in the rural area, and 60 per cent of the total population was aged under 20 years. Rwanda was [and remains] the most densely populated country in Africa with 273 people per sq.km. It had a low *per capita* income of US\$237, making it one of the world's 42 least developed countries. Then, as now, coffee and tea exports accounted for the bulk of country's source of income [4].

In May 1996, the new government set in motion a process to establish an e-government system in fulfillment of a commitment made by African governments under the Africa Information Society Initiative (AISI); "a visionary statement" on how Africa could use information and communications technologies (ICT) to accelerate economic and social development. The statement was also conceived, elaborated, and politically endorsed by Africans as a partnership with others committed to the AISI goals. Thus, while it was framed as a policy, AISI was also a consensual new way of doing things, that is a norm for how to modernize, and the basis for engagement with the international community. Based on this understanding, all African governments agreed, as a first step, to establish national information and communications infrastructure plans (NICIs), the national policy statements to domesticate these regional norms. By June 2005, Rwanda had surpassed this minimal target by far, becoming a model on information technology deployment in Africa. This development was due to a convergence of four key circumstances – the contribution of emigrants and refugee returnees, networking with

strategic epistemic communities, an activist political leadership, and an under-contested policy environment.

4.1 Emigrants and Refugee Returnees

By the time Rwanda made a commitment to implement the AISI framework in May 1996, the seed of integrating Information Technology (IT) into the country's development plan had already been sown. In the dying days of the 1994 civil war, Rwanda's incoming leadership was aware that more than a change of government was needed to resolve the more than half-century of political turmoil the country had undergone. Of particular interest was a long-term economic solution that would ease competition over natural (land) and state resources. In a predominantly peasant community whose economy is agriculture competition over land was particularly problematic. This was exacerbated by tangled up property ownership records following three waves of refugees exodus and repatriations in 1959, 1973 and most recently, 1994/5. With each refugee wave, those left behind appropriated property, whose ownership was then sanctioned by the incoming regime. This has left behind a messy record of property ownership that is yet to be resolved. The need to reduce dependence on land by transforming the human resource into the backbone of the country's economy partly explains how consensus was achieved and in record time (two years) over such a fundamental national policy and with support from politicians and stakeholders alike. But this was not where the story of e-government diffusion began.

The bulk of Rwandese emigrants involved in this process were refugee returnees who had been accustomed to an e-government norm in their countries of exile. They had also observed the economic benefits that accrued to countries such as Singapore under the "Intelligent Cities" program. In nearly all the interviews, my inquiry for a description of the vision of this much-touted 2020 "knowledge society", elicited the same immediate response, "to be like Singapore". The similarities between Singapore and Rwanda are not self-evident but were "common knowledge" among Rwanda's e-government entrepreneurs.

Many interviewees recalled that Rwanda resembled Singapore in its structural conditions therefore Singapore's success suggested Rwanda too had great growth potential. Both countries have small populations and spatial dimensions that exert pressure on their limited natural resources. But these conditions also presented opportunities. They meant fewer inputs would be needed to develop the country's human resource and infrastructure. Singapore had demonstrated that investments in IT education had a rapid rate of return on investment through outsourcing service jobs. Service-oriented jobs had the potential to get people off the land, and in turn, possibly ease up the political tensions associated with the tangled up private property history. Moreover, Rwanda's location in the middle of the African continent, as well as its multilingual status (Kiswahili/French/English), could be exploited to enable the country to serve as Africa's regional communications hub, just as Singapore's strategic positioning had made it a communications hub for its region.

While interviewees often drew parallels with Singapore in terms of its similarity to Rwanda's structural constraints and how they could be overcome through developing the human resource, for inspiration they looked to South Korea's successful transition from a low- to upper-income economy in three decades. *Prima facie*, these self-comparisons with Singaporeans and South

Koreans are only instructive on how to transform their economy. But the impact was deeper. These images enabled the Rwandese leadership to cease viewing the Rwandese people as a problem, but rather, as its greatest asset and resource for future development. Later, they propagated a view of Rwandese as a resource with the ability not only to create job opportunities in the domestic environment, but also become a net "exporter of the IT human resource to the Eastern African sub-region". It also freed the leadership from the sole pursuit of a legal route to resolve the prevailing property chaos, and be open to possibilities with the potential to render greater value to assets other than "fixed" and natural resources. Thus, it altered Rwandese self-understanding of their country as one lacking in assets for economic growth, into a country with great potential of becoming a middle-income economy by the year 2020.

But to get to this point, the returnees had to involve the political leadership in the construction of this vision of reality. To this end, they provided training informally to key political leaders on how to use IT. Once the leaders were sufficiently versed with the workings and use of the technology, they became part of the constructors and developers of this reality, and behind the scenes prodded open doors that were shut. Thus, when monopoly Rwanda Telecommunications, the sole regulator of the airwaves, resisted the licensing of Kigali Institute of Science and Technology and the National University of Rwanda as alternative Internet service providers, highly placed political leaders involved in the construction of this new reality exerted political pressure to create an opening. Eventually, the head of Rwandatel was replaced.

Emigrants and refugee returnees also became early contributors to Rwanda's human resource development and investment. Until 1994, Rwanda had never been a front-runner in the use of IT. In fact, by the time the civil war broke out in 1993, there was only one national public radio station and no television. Therefore, to transition into an information society Rwanda needed a massive injection of a human resource with the necessary skills. It came in two forms. Refugee returnees, emigrants and other Rwandese living in the diaspora became the early investors. Absent a developed private sector, Rwandese repatriated some of their foreign investments and set up local businesses. Later, professional Rwandese returnees who initially worked in the bureaucracy turned into business entrepreneurs involved in the IT industry. These ventures attracted new Rwandese professionals and expatriates as well.

Thus, what started off as a small circle of elites constructing a new vision of, and reality for, Rwanda drew in more people progressively in an iterative manner. Each cycle was greater than the preceding one; a cascade recursive pattern of e-government norm diffusion, in which output from one level of vision became the input into the next and into which a larger number of Rwandese were brought in. The more people became a part of this vision, the more persuasive it got, to the extent of drawing in other actors, including the international community.

The second source of human resource investment was the refugee returnees that already had IT expertise. Rwanda's refugee returnees differed in significant ways from other refugees in Africa. Rwanda's 1959 and 1973 refugee crises occurred at a time when the predominant refugee management approach based on

the Organization of African Unity's 1967 Convention was integrationist. It viewed refugees as a development resource and enabled their integration into the local communities, naturalization if they so wished, and repatriation only on a voluntary basis [3]. Modern day refugee camps that offer only minimal human resource development skills and opportunities have only impacted the last (1994/5) refugee exodus from Rwanda. This ensured that a majority of the refugee returnees from the first two waves had fully developed human resource skills equal to any available on the Continent, as they had benefited from opportunities equal to those of other Africans.

Rwandese refugee returnees therefore came from all corners of the world – developed and developing countries alike, even as far as China. Their professional exposure while in exile exceeded that of the average Rwandese who had remained in the country all the while. This human resource inflow was a boon to the new government. Involuntary exiles tend to be fiercely patriotic, thus Rwandese refugees gave up lucrative careers elsewhere to rebuild their country and lives. This patriotism combined with highly skilled refugee returnees that were already versed with e-government enabled a rapid diffusion and acceptance of an IT-based economic structure, adoption of the e-government norm, and its diffusion. The Rwandese refugee returnees resemble those of countries such as Uganda whose refugee crises also occurred before 1984 whose returnees brought back a fully developed human capital. They differ from those of the Democratic Republic of the Congo, Sierra Leone, and Burundi, who generally flee to neighboring countries and are confined to camps with limited human capacity development. As a result, returnees are often less able to contribute to their countries economic development in the short term.

A major priority of the government after the civil war was to rebuild government institutions and systems that were razed to the ground during the war. Because the construction of this new norm was taking place when resources were being mobilized to construct and rehabilitate government buildings, government agencies were able to factor into their resource mobilization strategies for financing an e-government system. Also, because many of the professional refugee returnees initially worked for the government, they were able to propagate the culture of e-government they had acquired abroad. All these features combined to minimize the arousal of bureaucratic politics that might have stifled the spread and acceptance of e-government.

The foregoing demonstrates that emigrants, especially refugee returnees, played an important role in the development of e-government in Rwanda in four ways. They conceptualized the initial vision of an information society as the basis for Rwandese economic growth. They called on powerful success images of Singapore and South Korea to facilitate the construction of this reality with the broader political community, and thereby transformed the self-perceptions of a critical group of actors. They injected into the system a substantial, badly needed human resource input in the aftermath of the genocide. And being highly skilled and socialized into e-government through exposure abroad, they supported its diffusion. Lastly, they provided financial investments to kick-start the development of an e-government system. Viewed over time, emigrants, especially the refugee returnees not only sustained this vision through its continued practice and creation of supportive institutions such as the NICI

plan, they also carried their vision with them into new sectors and areas of activity.

4.2 Networking with Epistemic Communities

Despite the zeal and commitment of returning refugees, Rwanda still faced major resource gaps. In particular the country lacked sufficient technical capacity to deal with the level of technical knowledge, skills and capabilities needed to plan and manage the type of socio-economic transformation the country aspired to. Although they had a broad base of potential users, they lacked network engineers, programmers and information and database managers that would render an e-government system functional. In developed countries, this expertise typically resides in private institutions such as academia and the private sector, therefore governments here were able to outsource it initially, before building in the institutional structures and in-house expertise. In contrast, a substantive number of domestic expertise with high-end IT skills was lacking in a majority of the African countries.

Rwanda turned to the IT epistemic (knowledge) community; that is, global networks of IT experts that they had engaged with during their careers abroad, and through regional intergovernmental organizations. These include the UN Task Force on Information and Communication Technologies, and the forum for African academics hosted by the United Nation's Economic Commission for Africa. This approach enabled the government to manage the high financial investment that would have accompanied alternatives such as private contracting. This networking approach also attracted individuals that were more sympathetic to the plight of post-genocide Rwanda, and whose primary motive was not profit. It is with the support of these communities that Rwanda developed its ICT-based Socio-Economic development Policy in 1997/8, as well as the first of four five-year NICI plans, and the 2020 Vision for Rwanda (V4R). The NICI Plan has eight sub-plans of which e-government is one of the most advanced.

Rwanda also started tapping into private initiatives that support IT development such as the CISCO Systems' Networking Academy Program. By drawing in private companies whose core expertise is information technology, the government was able to tap into first-rate private companies' technical knowledge without paying actual market rates. Where networking was insufficient the government received financial assistance from foreign governments and entities, to hire a few external experts who were also linked to these, and their own, epistemic communities. The involvement of Cisco Systems in the development of a regional center based in Rwanda has contributed to the propagation of the vision and development of a critical mass of technical engineers that are able to support e-government diffusion. The center is a training outlet in network design and management for high school graduates who are unable to pursue undergraduate studies. With the CISCO regional center located here, and Rwanda's recent membership into the Eastern African Community as well as its multilingual status, it is foreseeable that Rwanda will soon be exporting technical expertise to the region.

But it is the social transfer of technology and technical skills through this strategic networking that is of interest here. Networking boosted Rwanda's human resource capacity. Staff attending workshops or meetings organized under these fora had an indirect avenue to learn and accumulate knowledge. In so

doing, Rwanda was able to fill in the conceptual gaps that would have hampered the validation of this process. Networking opportunities were especially useful for computer engineers and staff running government IT directorates because many of these were among the first graduates of Rwanda's National University, and lacked extensive experience in their fields. Because epistemic communities possess a diverse range of expertise, involvement with such communities enabled a few key Rwandese to deepen their understanding of what an IT-based economy would require and the gaps in the Rwandese system, and to then bring these ideas back home. Individuals from the region's and international epistemic communities were often invited to Rwanda's national planning processes. This facilitated the engagement of a larger group of Rwandese with a few key external actors over issues that would advance Rwanda's effective construction of new norms, and its identity.

These two processes, networking with epistemic communities from the voluntary and commercial sectors facilitated both a socially constructed, and market-based, transfer of technology. Besides its concessional value, there is a lot to learn from the social transfer of technology. Progress and project evaluation reports suggest that the major weakness in Rwanda's e-government remains its human resource, and emphasize the need for more training for this workforce.

But what kind of training? What is the value of social technology/technical transfer compared to other formal and market-based transfers? What in particular are its contributions? What are the diffusion mechanisms through which it occurs? In what ways do they constrain or open up possibilities for overcoming structural constraints to e-government diffusion? These constitute some of the questions requiring further research.

They are useful because I found that part of the dilemma for the government was a desire to have students graduating from its institutions taking charge of the system, and as fast as possible. Yet, many of these IT professionals are not as well linked to epistemic communities as are the more experienced professionals and those with foreign experience. Mostly due to a lack of this broader exposure, the e-government systems managed by the new graduates were less developed and not as well conceptualized as a part of the broader institutional fabric.

4.3 Political Leadership

Perhaps more than any other factor cited is the assumption that in countries with little access political leaders are necessary for the diffusion of e-government and IT more generally. But it is not always clear how leadership matters. Rwanda provided a moment to begin an exploration of this issue.

The role of political leaders in promoting e-government and the development of IT is perhaps what sets Rwanda apart from other African countries. Nearly all interviewees recounted a story of e-government that started and ended with the country's President, Paul Kagame. Kagame was the head of the rebel Rwandese Patriotic Front that took over the country's leadership at the end of the 1994 civil war. In the new government he headed the defence ministry, until he was elected President in the country's first democratic elections in 1997.

All the interviewees almost always referred to Kagame as the ICT "champion", and was often credited as the originator of the idea that the country's economic growth could be IT-driven. He had embraced it and pursued it to the extent that many erroneously attributed its origin to him. These assumptions are correct to the extent that although Kagame was not the source of the vision, he was in the initial small core of returnees considering whether an IT-led economic structure in Rwanda was even possible. Since becoming President and as the number of those engaged in discussions about an IT-based social revolution broadened, Kagame's direct involvement decreased. He appointed a personal advisor on ICT instead, to keep him up-to-date on IT.

The President's role as champion of ICT emerged very early, almost shortly after his rebel Rwandese Patriotic Front took over, thanks to fortuitous co-incidences. Then vice-President and head of the defence portfolio, Kagame had several staff members working under him and others in strategic ministries that had previously worked with international organizations and international development agencies such as the World Bank and United Nations Development Program where they were exposed to debates about the potential for Africa to "leapfrog" development through the use of Information Technology. These bureaucrats approached Kagame and over a period of time exposed him to these debates, including walking him and others through the practical aspects of IT. These developments coincided with the new Rwandese leaders' search for long-term economic solutions.

Although Kagame caught onto the vision quite early, steps towards a social revolution did not unfold fully until after he assumed leadership as President. In this new position, he appointed an independent task force to design the national development plan with information technology as the cornerstone. It is here where the ideas that had been part of a small core were first exposed to a broader audience. The task force engaged directly with each other and occasionally, formally and informally with others outside the circle through "consultations". These consultations can be understood as an attempt to broaden and review the norms and structures conceptualized by the initial small core group. The task force then drafted the national plan, which was later deliberated upon and approved by a national stakeholder workshop in 2000, and again in 2002.

Still, there were a few critical actors, politicians in particular, that chose to dissociate from the process. Kagame became the one to mobilize their support in the hope that all members of the Transitional National Assembly – the Senate and House – would be involved in the ICT agenda because they approved the budgets and some served as political heads of the ministries as well. Thus, prior to the approval of the country's NICI plan, the entire legislature participated in a half-day debate session chaired by the President on the role of ICTs in the country's economy. The legislature subsequently approved the plan unanimously, and proceeded to set up the first and most comprehensive administrative electronic system for themselves. To entrench this e-government practice and norm, members of the Transitional National Assembly underwent several training sessions, and the Assembly became one of the first institutions to have a fully operational e-government system.

This scenario speaks to the first role of political leadership in the rapid development of e-government in Rwanda, and social diffusion of technology more generally. The diffusion of new norms is particularly challenging where two conditions obtain: they are alien to the host environment, and the requisite structural and institutional supports are lacking. This is because diffusion demands a double investment; first in the dialogical construction of the guiding norms and rules through the harmonization of members' self-understandings, and then setting up the structural framework. There is no guarantee that consensus will emerge regarding the nature of investment in the process. Political leadership, through the authority of the office itself, is invoked as necessary to bring into the fold dissenting voices. Thus, whereas political leadership can encourage the construction of reality, it has the potential to constrain it. The diffusion of e-government by different agencies displayed a similar logic. Evaluation reports indicate that where the ministry's leadership was committed, e-government adoption had progressed rapidly. Moreover, the ministries' ICT team leaders had even taken the proactive role of mobilizing external resources independent of the government, and that the recommended five per cent allocation of the agency's annual budget to ICT directorates was implemented consistently.

Kagame's second role as leader is publicizing Rwanda's new image and aspirations to the domestic and international community. He never passed up opportunities – in fact, he was proactive – in all international initiatives. He was one of only two African Presidents to attend the African regional preparatory meeting for the World Summit on the Information Society (WSIS) in February 2005, and also led his country's delegation to the first and second WSIS in 2003 and 2005 respectively. Kagame's self-identification with this new Rwanda was so strong that it had become his number one priority. One speaker at the National Workshop reported that the new World Bank President Paul Wolfowitz was surprised during his visit to Rwanda that Kagame's sole request to the Bank was financial support for the country's ICT agenda. Presumably, few, if any, African leaders have demonstrated comparable levels of commitment or even offered minimalist visions of the same whether to the domestic or international publics.

But why is Kagame referred to as "ICT champion"? According to interviewees, a champion is an individual who is a winner, and perceives himself as such. It is this attitude, interviewees said, that persuaded even initially skeptical and apathetic political leaders that economic growth through ICTs is possible and the way of the future. In other words, Kagame's self-identification with the ICT cause was sufficiently convincing for potential domestic opponents to sign on, and to reassure proponents. By personally engaging with the vision of a Rwandese information society, he was able to identify with it and assign a high priority to its development. Also, he became the country's embodiment of the new Rwanda, and his commitment to the vision has enabled him to mobilize substantial financial resources for the process. Further, the shift by all members of the Transitional National Assembly to an electronic conduct of House and Senate business is likely to spill over to members' local offices and interactions with other public agencies, and thereby broaden and entrench the system. Thus, at the regional and local levels, members of the Assembly are likely to replicate Kagame's role, but this constitutes another area for potential research.

What is evident from Kagame's leadership role is, again, a cascade recursion pattern in the construction of Rwanda's new identity and diffusion of the e-government norm. Equally fascinating for scholars on leadership is the question of the (re)construction of a leaders' identity upon assuming of office.

4.4 An Under-Contested Policy Environment

One of the most surprising findings of the study, and which was critical to the rapid deployment of e-government, including the development and implementation of the Vision, and NICI plan was an under-contested policy environment.

Rwanda as a state was established in 1885, and gained independence from Belgium in 1962. But the country is only a decade old. The interviews and informal conversations I had even with ordinary people on the evolution of e-government were punctuated by "after the war" (1994). Apparently, the development of e-government, and ICT in general, is part of the social construction of the state, to give it a new history and identity. This was possible due to the combination of civil war and, tragically, genocide. The civil war destroyed preexisting political institutions, not a rarity in Africa. What is unique is the extent to which the new government had free reign in their reestablishment, to the extent of setting up a bi-cameral national assembly akin to that of the US, a rare political arrangement in Africa. This was facilitated not so much by the civil war, as the genocide, which had claimed so many people that it demolished pre-existing interest groups. It also silenced the international community whose failure to avert the genocide robbed it of any moral authority to impose the types of conditionalities that accompany foreign aid, and foreclosed an avenue through which current Rwandese exiles might have found their entry.

Timing was key. Kagame, the chief e-government proponent, assumed the country's political leadership shortly after the war. But by 2005, just over a decade after the war, hardly any civic advocacy groups within the country had organized or re-grouped around policy issues. Moreover, of the large number of domestic and transnational non-state actors operating in the country, hardly any were focusing on development issues, which would have drawn them into the IT agenda. A couple were focusing on policy issues, but the majority was still involved in humanitarian work and state reconstruction. There is a nascent private sector, mostly in the form of small entrepreneurs. And with hardly any foreign investors flowing into the country, the government has been proactive in supporting entrepreneurs, especially those aligned with the long-term vision of an IT focus, through partnerships and outsourcing of e-government services. But even in the agriculture business community that is likely to lose the most from this development, there is hardly any political activism.

As a result, the policy environment is devoid of organized interest-group politics that might have contested efforts to establish, or particular aspects of, the e-government system. Having been socialized over the last eight years into a new way of work, and given the disruption to the bureaucracy in the aftermath of the genocide, e-government bureaucratic politics was almost non-existent. It reared its head early on through Rwandatel but was quashed through political pressure, and subdued. The main source of opposition, Rwandese refugees and exiles, lacks a visible and effective foothold domestically or in the international community. It is this brief window of opportunity providing an

undercontested policy environment that has enabled Rwanda to make greater progress than many of the African countries.

In the absence of serious domestic political contestation, a small albeit revolutionary idea was presented, developed, and propagated into the broader community without insurmountable opposition. At the same time, the government has been able to mobilize much needed financial resources through the emigrant, refugee returnee, and international communities, and to deploy these resources without donor impositions that constrain other countries. Moreover, Rwanda's development framework and e-government initiatives are premised on a neoliberal agenda, which minimizes further opportunities for foreign interference. It is this under-contested policy environment combined with a President who is the chief proponent of this new reality that set Rwanda apart from other developing, and sub-Saharan African countries' pursuit of e-government.

5 IMPLICATIONS

This discussion leads to four questions. Were the four conditions sufficient to propel Rwanda to a position of e-government leadership in Africa? How do these conditions help us understand the unfolding story of e-government in Africa? What are the implications of these findings for theory? Lastly, what potential areas for further research on e-government in Africa emerge?

The least significant component in this mix is networking with strategic epistemic communities. This is because the expertise that they provided could have been acquired through alternative means such as technical assistance, internships, sabbaticals or even hiring of expatriates, some of which the government had already done. Still, it was useful in advancing the process faster than might otherwise have occurred, and at a lower cost. It also enabled Rwanda to narrow the human resource gap it lacked compared to other African countries. Although Rwanda was generating a critical mass of IT experts, other African countries are better placed with regard to the diversity in the levels, and areas, of their resident human resource expertise.

Emigrants, especially refugee returnees, were a critical link to the initiation of e-government in Rwanda. They provided the ideas, as well as financial and human resources required to overcome serious structural constraints. They also provided a useful link to diverse epistemic communities located outside Rwanda to bridge a crucial knowledge gap. Whereas emigrants were a defining factor for Rwanda specifically due to the impact of genocide, and in their visioning capabilities, by themselves they would not have made much difference. Their vision of an IT-led socio-economic transformation remained incubated until a leader emerged to carry it forward. Moreover, once the idea began to take root, even low-intensity bureaucratic politics from Rwandatel was sufficient to frustrate it. Finally, without a national plan it is unlikely even magnanimous emigrants would have willingly repatriated their funds for investment. Thus, while necessary, the contribution of emigrants and refugee returnees was not a sufficient condition for effective e-government diffusion.

It is the remaining two conditions – political leadership and an undercontested policy environment – that were the most important in Rwanda's achievements.

President Kagame's commitment to the Vision, coupled with an ability to carry with him other leaders at the domestic level and to persuade the international community that the vision was attainable have had great impact. The vision has penetrated beyond the bureaucracy, into important national institutions, including the formal education system from the grassroots level in primary schools to private and public institutions of higher learning, and other vocational schools. All college graduates are required to have basic training in computer skills and IT. The use of electronic technology in the village-level trials of those who committed genocide is exposing even illiterate people to a culture of high-technology use that is lacking in many African countries. These attempts to demystify technology are likely to elicit conversations and engagement about IT, as will the use of information kiosks for public use that are being set up in public offices. In short, Kagame has used his office effectively to persuade resistant, skeptical and apathetic political leaders, to mobilize support from outsiders and undercut voices of opponents, and by embracing the idea, to mobilize the Rwandese society to start reforming strategic social structures and to envision a new Rwanda. There are few leaders from sub-Saharan Africa who match this.

But Kagame's success is not without reason. He benefited from an undercontested policy environment that enabled him to engage in the construction of the country's vision and policy through a dialogue that was heavily top-down. Rwanda was peculiar in two regards. First, it lacked the kind of contestation that has characterized many of the other sub-Saharan African countries. And second, this top-down planning process made for a rapid turnaround in defining a fundamentally political issue – the transformation of the economic structure – without serious political ramifications. Both approaches to planning are unlikely to proceed as quietly in a majority of sub-Saharan African countries.

In a recent comment on Kenya's e-government process, the immediate former Minister for Economic Planning Anyang' Nyong'o advanced two factors for the long duration it had taken to develop an e-government strategy, and now to implement it. First, being held hostage to history. He suggested that the culture of fear, monopoly and control that had characterized the IT sector was difficult to let go. A second challenge is the problematic nature of pursuing democracy and economic reforms in tandem [16]. In a democratic environment, interest groups with unscrupulous aims could bog down legitimate and progressive reforms. The culture of "participatory development" that requires broad based consultation with all interest groups prior to the design of public policy has become a part of Africa's social reality, and is partly why e-government diffusion has taken an incremental approach. Resource reallocation, the institutions of priority, and opposition by potential losers all necessitate dialogue to reconstruct new rules, norms and institutions among diverse players, sometimes over intractable issues.

Because many of the stable African (and other developing) countries now tend to resolve disagreement through dialogue, their policy environment is likely to be fiercely contested, and thus slower at accepting and molding foreign e-government norms to fit their environment. It is this history that Rwanda was freed from briefly by genocide, and enabled it to construct new norms.

The ability of Rwanda's leadership to maximize on this window of opportunity that opened up to them made the difference.

The foregoing suggests that by themselves, none of the four conditions outlined above in Rwanda, would have assured the rapid diffusion of e-government. Rather, it was their convergence. But what are the implications of these outcomes for theories of technology diffusion?

A constructivist understanding of e-government diffusion renders explicit the difference between e-government diffusion in developed and developing countries, and why the international community's efforts to set up structural and institutional structures to bridge the capacity (diffusion), but not adoption, divide are insufficient.

Because e-government was a latecomer in developed countries being a spill over from e-commerce, the delay was advantageous. The socialization of a large population in advance of its introduction, as well as the eventual demand-driven process that emerged greatly diffused the potential for group politics. Because a large proportion of the society not only engaged in this practice but was already locked into it, e-government diffusion became more a question of overcoming bureaucratic, and less interest group, politics. Other developed countries where the market rules, institutions and practices resembled those of the US were also able to adopt and adapt to the norm equally first and extensively, and perhaps more efficiently by learning from the US' mistakes.

But where these practices, rules, norms and institutions are absent, that is in developing countries, e-government requires greater effort to domesticate, and has had to take the form of a policy. Because new policies elicit politics, the introduction of e-government in developing countries has not been smooth. Yet these political factors have remained largely ignored with the exception of the politics associated with the deregulation of the telecommunications sectors. The foregoing discussion suggests that absent sudden unexpected occurrences, e-government diffusion processes are likely to be slower where democratic practice is more mature because privileged interests groups will resist a change of the existing social practices, rules and norms. Therefore, models of e-government diffusion in developing countries that fail to integrate the behavior of agents in the system, that is, political variables associated with group relations, only tell part of the story.

Although this study provides insights about the patterns of e-government evident in Africa, it leaves many interesting questions unanswered. There is need for a comparative study of donor and other philanthropic contributions towards e-government initiatives in sub-Saharan African countries to illuminate just how important resources were in this mix for Rwanda, compared to other countries. Exploratory case studies of countries such as Zambia where e-government appears to have bottomed out in a country that was among the first to adopt IT might also provide insights about the significance for sub-Saharan Africa of the four conditions stipulated in this study. A comprehensive understanding of e-government calls for an examination of both the diffusion and adoption dimensions of e-government. Thus a similar study of Rwanda focusing on the adoption dimension would offer greater clarity of these dynamics. Finally, similar studies may be undertaken in other regions to determine whether

other forms of social interaction exist in the e-government diffusion processes, whether particular forms of social interaction are suited to particular political environments, and finally, within the context of leadership scholarship, whether and if so how, leaders in transitional regimes re-construct themselves and their societies.

6 CONCLUSION

Compared to many developing countries, particularly those in sub-Saharan Africa, Rwanda's progress in e-government diffusion has been phenomenal, and calls attention to a pattern of e-government diffusion in the region that appears to defy causal explanation. This study established that part of this messiness was a result of agential intervention, specifically interest group and bureaucratic politics. Designed as an exploratory case study, Rwanda was selected in order to understand why it had achieved more progress than would have been expected given its structural conditions, and to depart from previous studies that focused on failed cases. Framed on the basis of social relational and group theories, this study found that the convergence of four factors enabled Rwanda to make greater advances than other countries with similar conditions. These conditions are the contribution of emigrants, including refugee returnees, an activist political leadership at the highest levels of government, networking with strategic epistemic communities, and the presence of a relatively under contested environment. It was this last condition, however, that was critical and enabled the convergence to happen.

This study also argued that the model of e-government diffusion in developed and developing countries differed substantially for meaningful comparison. It also suggested that the indicators used to measure e-government diffusion are misleading. They are premised on a modernization framework that uses measures derived from the realities of the country "most advanced" along the dimension under consideration. In so doing, they fail to capture innovations taking place in other countries that make up for the social and structural gaps, and thus under-estimate their achievements.

Finally, this study highlighted potential areas for further research such as the significance of donor funding in Rwanda's forward leap, and forms of social interaction associated with e-government diffusion in other regions and at varying levels of development.

7 APPENDIX I: RESEARCH METHODS

The data collection for this exploratory case study involved 24 in-depth interviews lasting between one to one-and-a-half hours with key informants drawn from government (7), private sector (7), academic (4), intergovernmental organizations (2), politicians (1) and non-governmental (3). Only three interviewees were women. I also participated as an observer at the two-day National Validation Workshop, a multi-stakeholder meeting convened to consider and approve the evaluation report of the first five-year implementation of the 2000-2005 national information and communications infrastructure plan. During the workshop, I also conducted informal conversations with some of the participants. Lastly, I used archival research to cull historical and secondary data on ICT development in general and e-government adoption more specifically from feasibility studies, and progress and evaluation reports carried out by funding agencies for internal reporting during the 2000-2005 NICI plan period.

The interview schedule consisted of nine key open-ended questions that served as conversation initiators and guides to focus interviewees. I employed probing techniques extensively to pursue interesting lines of inquiry that emerged from the interview. One of the shortcomings of this method, however, was the inability to return to prior interviewees to verify data that emerged in later interviews. The contribution by emigrants was one such issue that only surfaced towards the end of the study. It failed to emerge early in the interviews in large part due to my own attempts to avoid opening up lines of inquiry that could lead to a discussion of the politically charged and psychologically sensitive issue of genocide, which I was ill equipped to handle nor ethically permitted to touch on. I realized too late that the role of emigrants could have been discussed without moving in that direction.

A combination of purposive and snowball sampling techniques was used to select the informants. I assumed that interviewing actors that had been actively engaged in the process from inception would provide me with rich and accurate data within a relatively short period of time. Second, given that the IT epistemic community in Africa is quite small, I suspected that the IT community in Rwanda would be even smaller, therefore most interviewees would likely have been interviewed already on these or related issues, and would likely have reflected in depth and on more than one occasion on the issues of interest to me. Therefore, they were likely to provide some consistent responses on many of the issues.

To minimize bias, I interviewed informants from public institutions where the greatest progress had been made, and those that made least progress in e-government diffusion. Further, the bulk of this data came from government ministries; only a couple of other public agencies were studied. The field research duration was June 26 to July 23, 2005.

8 ACKNOWLEDGEMENTS

I thank the Department of Political Science, the Moynihan Institute of the Global Affairs, and the Maxwell School Dean's Office for their financial support to carry out the field work, my host institution Rwanda Information and Technology Authority, Professors Stuart Thorson, Audie Klotz, Jon Gant, David Richardson, the 2005-06 Maxwell School's Goekjian Scholars, and two anonymous student peer-reviewers for their comments on various drafts. Errors of substance remain my responsibility.

9 REFERENCES

- [1] AISI. Status Graph, NICI Graph, UN Economic Commission for Africa, Addis Ababa, October 1995.
- [2] Choucri, N., Maugis, V., Madnick, S. and Siegel, M. Global e-Readiness - for WHAT? *e-Business@MIT*, Massachusetts, 2003.
- [3] Crisp, J. Africa's Refugees: patterns, problems and policy challenges. *Journal of Contemporary African Studies*, 18 (2). 157-178.
- [4] Esselaar, P., Hesselmark, O., James, T. and Miller, J. A Country ICT Survey for Rwanda. Final Report. Swedish International Development Agency, S. ed., Esselaar and associates, Kigali, Rwanda, 2001.
- [5] GOR, Policy and Strategy for Rwanda. in *National Workshop on Information and Communications Technologies, November 30 - December 2, 1998.*, (Kigali, Rwanda, 1998), Government of Rwanda.
- [6] Heeks, R. e-Government in Africa: Promise and Practice *Institute for Development Policy and Management, iGovernment Working Paper Series.*, Manchester, UK, 2002.
- [7] Jensen, M., Information and Communication Technologies (ICTs) in Africa - A Status Report. in *Third Task Force Meeting*, (United Nations Headquarters, 2002), UNICTTF.
- [8] Jensen, M., The Status of African Information Infrastructure. in *First Meeting of the Committee on Development Information (CODI)*, (Addis Ababa, Ethiopia, 1999), UN Economic Commission for Africa.
- [9] Jensen, M., Toward an African Information Society: Lessons Learned in Harnessing New Information and Communication Technologies (ICTs) for Development in Africa. in *Information and Communication Technology and Development: RAWOO Lectures and Seminars*, (The Hague, The Netherlands, 1998), RAWOO, 51-58.
- [10] Katz, E. Review Essay. Theorizing Diffusion: Tarde and Sorokin Revisited. in Lopes Paul and Mary Dursee, s.e. ed. *The Social Diffusion of Ideas and Things, Annals of American Academy of the Political and Social Sciences*, Sage Publications, New Delhi, 1000 Oaks, London, New Delhi, 1999, 144-155.
- [11] Kinyungu, C. Kenyan MPs humbled in Rwanda *The Daily Nation*, Nairobi, 2005.
- [12] Mbarika, V., Jensen, M. and Meso, P. Cyberspace across sub-Saharan Africa. *Communications of the ACM*, 45 (12). 17-21.
- [13] Milner, H., The Diffusion of the Internet Globally: The Role of Political Institutions. in *2003 Annual Meeting of the American Political Science Association*, (Philadelphia, Pennsylvania, 2003), APSA.
- [14] Nations, U. Global E-Government Readiness Report 2005, From E-Government to E-Inclusion, UN Department of Public Administration, New York, 2005, 270.
- [15] Norris, P. *Digital divide*. Cambridge University Press, Cambridge, 2001.
- [16] Nyong'o, A. Planning for Policy-making and implementation in Kenya: Problems and prospects. in Florence, E. and Eder, L. eds. *At the Crossroads: ICT Policymaking in East Africa*, East African Publishers/IDRC, Nairobi, Kenya, 2005.
- [17] Policy, P.C.o.I., Roadmap for E-government in the Developing World. 10 Questions E-Government Leaders Should Ask Themselves. in *The Working Group on E-government in the Developing Countries*, (2002), University of Southern California.
- [18] Rogers, E. *Diffusion of Innovations*. Free Press, New York, 1995.
- [19] Rugondihene, J., Implementing the NICI Plan 2005: Reviewing the performances of the implementation agencies. in *NICI Plan (2001-2005) Validation Workshop, Kigali, Rwanda, 27-28 June 2005.*, (Kigali, Rwanda, 2005), RITA.

e-Governance in Africa, from theory to action: a practical-oriented research and case studies on ICTs for Local Governance

Gianluca Carlo Misuraca

Executive Master in e-Governance,
Chair MIR, College of Management of
Technology, Ecole Polytechnique
Fédérale de Lausanne – EPFL
EPFL-CDM-MIR, Odyssea, Station 5,
CH 1015, Lausanne, Switzerland
+41 21 693 00 11

gianluca.misuraca@epfl.ch

ABSTRACT

The paper is an extract of the Research on “e-Governance in Africa: from theory to action: a practical-oriented research and case studies on ICTs for Local Governance”, that is being conducted by the author within the framework of the “Executive Master in e-Governance” 2004/05, at the Ecole Polytechnique Fédérale de Lausanne – EPFL. The research focuses on the context, theory and thinking around the issue of ICTs and local governance, particularly in Africa. After briefly discussing the basic concepts, from government to governance, and the role of local level authorities, presenting the benefits and limits of introducing ICTs in government operations, the paper identifies the common elements for providing proposed definitions. In this connection, discussing the paradigm shift from e-Government to e-Governance, it elaborates on the potential of ICTs at local level (“e-Local Governance”). Furthermore, the paper presents the results of four case studies conducted evaluating selected projects in Senegal, Ghana, South Africa and Uganda. Based upon the findings of the case studies, the paper presents the conclusions and some recommendation on the way forward to make effective use of ICTs for better governance and promoting economic development at local level.

General Terms

Management, and Measurement

Keywords

e-Governance, e-Government, ICTs, local governance

1. INTRODUCTION

Decentralisation and locally-controlled administration are increasingly recognized as basic components of democratic governance and provide an enabling environment in which decision-making and service delivery can be brought closer to the people, especially the poor and the marginalised. Community participation in decision-making, planning, implementation and monitoring and backed by appropriate institutions and resources; along with effective decentralisation can, through ensuring greater accountability, responsiveness and participation, result in local services that are more efficient, equitable, sustainable and cost-effective. The integration of Information and Communication

Technologies (ICTs) in the governance processes can greatly enhance the delivery of public services to all citizens and thus the overall objective of improving the performance of governance systems at all levels, as well as transforming the relations among the different stakeholders involved in the governance system, and thus influencing the policy-making process and the regulatory framework. The potential of ICTs for Governance in developing countries, however, remains largely unexploited and local governance is, in general, given little attention within national ICTs and e-Government strategies. However, the broad assumption that decentralisation policies can influence good local governance and that the use of ICTs can greatly increase this influence has yet to be proven. To date there is little empirical evidence of the “multidimensional” effects of ICTs on local governance which can, in turn, inform national e-Governance policies.

2. OBJECTIVES

While there are some examples of linkages between ICTs and local governments, the causal connection between ICTs and innovation in local governance for socio-economic development is little understood. Also, the recognition of the potential of ICTs for local governance comes from a few successful pilot applications around the world. While attempts are currently underway to critically evaluate some of these projects so that the real extent of their impact can be understood and the factors inhibiting impact can be identified, as yet there is no common assessment framework from which lessons can be drawn in examining the link between ICTs and local good governance, especially in Africa. The research, drawing on an “institutionalist” framework, investigates how, on the one hand, ICTs affect policy outcomes within a given institutional setting and, on the other hand, what the key drivers or factors of success are in implementing strategies and programmes that include integrating ICTs at Local Governance level.

3. METHODOLOGY

The theoretical framework underpinning the research assumes not only a direct intervening effect of ICTs on policy outcomes, but a possible indirect effect of ICTs on the institutional settings themselves.

To answer the research questions mentioned above, the research design uses different levels of analysis and perspectives, based on practical-oriented research at the country level, where analysis focuses on four selected projects being implemented in Africa as case studies. This will begin creating an environment of lessons to be learned from each other and to inform about the factors of success in relation to enhancing governance and reinforcing democracy using ICTs.

The case studies are based upon “field-missions” undertaken within the framework of the Acacia and Connectivity Africa Dissemination Activities of the International Development Research Centre (IDRC), and on behalf of the United Nations Economic Commission for Africa (UNECA), in preparation of the Second Phase of the World Summit on Information Society – WSIS-II (Tunis 2005).

By investigating the relationships between ICTs and local governance, the paper provides empirical evidence of the dynamics, outcomes and implications for policy and practice of the integration of ICTs in local governance systems in Africa; also, key drivers for effective integration of ICTs into local governance systems are identified.

4. CONCEPTUAL FRAMEWORK

While it is not the scope of this paper to enter into detailed discussion of basic concepts or providing definitions of terms, a review of international literature and especially the background papers produced within the context of the Global Forum on Reinventing Government and the UN World Public Sector Reports, provides various definitions and arguments about the basic concepts around the theory of the transformation of the State, or what can be defined as the “paradigm shift” from government to governance[1] [2]. In fact, while Government and Governance both have the same root word, governance is about more than just government. It is a complex yet universal force that exists in all societies. [3]. The word governance has its origin in the Greek language and it refers to steering [4]. As an act of steering a people’s development, Governance is about processes, not about ends. While the study of “Government” is primarily concerned with understanding the institutional means through which public management is realised, “Governance” is concerned with the broader relationships between citizens and those institutions [5]. These matters are brought into additional relief when seen through the prism of ICT applications, where e-Government is concerned with the service delivery and transactions undertaken by institutions in support of the variety of public management activities, while e-Governance is more broadly concerned with the outcomes that may be enabled through the use of ICTs to support public involvement in public management. As a concept and in practice, e-Governance seeks to realise processes and structures for harnessing the potentialities of ICTs for the inclusion of citizens in the democratic processes of public sector management, service design and delivery and towards achieving good governance [6]. On a point further of conceptualization, a “working and evolutionary” definition of e-Governance can be: *“the use of ICTs, and especially the Internet, to adopt a new conception and attitude of governing and managing where participation and efficiency are required of all the partners linked in a network”* [7]. e-Governance is therefore a new way of co-ordinating, planning, formulating and

implementing decisions and operations related to governance problems, using ICTs as a medium of communication and partnership-development. Governments can utilise e-Governance to re-invent themselves, get closer to the citizenry and forge closer alliances and partnerships with diverse communities of interest, practice, expertise, conviction and inter-dependence within the context of local, national and international development agendas. The terms e-Government and e-Governance, therefore, are not interchangeable, since they refer to two different concepts. e-Government, refers to the use of ICTs as a “facilitator”, through “reshaping” the role of Governments, providing tools to support public service reforms, enhance public administration management and public sector performance vis-à-vis the private sector and citizens. It describes the active use of ICTs for the purpose of public service delivery only. Thus it is a narrower term than e-Governance which may be understood as the use of the electronic medium to facilitate an efficient, speedy and transparent process of disseminating information to the public and other agencies, and for performing government administration activities. e-Governance is generally considered as a wider concept than e-Government, since it can bring about a change in how citizens relate to governments and to each other. It is also about moving beyond passive information-giving to active citizen involvement in the decision-making process [8].

The relationships between ICTs and governance are in fact multiple. However, it should be underlined that, when discussing the integration of ICTs in administration, the focus is on the promotion of Governance using ICTs as a tool, rather than the ICTs as an end in themselves [9]. e-Governance is a growing phenomenon within public sector institutions around the world and is emerging as a significant discipline within the field of public administration and management in general. But understanding and managing e-Governance (as a concept and in practice) is a challenging and sometimes unwieldy task [10]. The concept of e-Governance is evolving and efforts to stabilise and clarify its operational implications must be made. In general terms, the debate regarding e-Governance is most often polarised between those who feel that ICTs will enhance the participation of citizens in the government policy decision-making process, and those who feel that it will simply be business as usual via a new medium [11]. Also it is argued that this will enhance the economic performance of developing countries, either through the creation/strengthening of the ICT industry, or through easier access of the private sector to crucial public sector information, and the increase of the “capacity for public service delivery of basic social services, public administration reform, integrated planning, increased citizen participation in decision-making, decentralisation, transparency, accountability and combating corruption” [12]. Within this theoretical and conceptual framework, what is to be underlined is that the possible impact of e-Governance can be divided into three main dimensions of policy objectives: the economic dimension, the social dimension and the political dimension, thus making it “multi-dimensional”. In addition, it has to be considered that, especially in developing countries, the public sector is increasingly seen as the main engine to bridge the digital divide at country level. Public agencies can start acting as model users of ICTs and be catalysts for others to follow. The public sector tends to be the biggest provider of local content and it can nurture and foster the further development of the local ICT industry. For this reason, the enhancement and/or

building of the capacity of public bodies and government agencies in the use of e-Government applications, promoting at the same time the accessibility of businesses and citizens to internet and government services on line (what can be called the e-Governance capacity), will improve knowledge through information availability. In other words, appropriate e-Governance initiatives can lead to strengthened conditions for good governance. Development of e-Governance is therefore not only a technical issue but also a political one. Here, though, we must recognise this high cost of failure, and look for ways to reduce risks. We should therefore understand the external and internal barriers to introducing ICTs in governance, and the needs and challenges for avoiding failures and implementing successful e-Governance. To do this, countries need to define priorities within the framework of their national policy goals, vision and strategic objectives and evaluate ICTs applications as they draw on scarce available resources and add different value to and impact on the governance process [13]. The development of such a policy framework cannot be done without considering the “local” component and the community development and how these integrate and use ICTs, in order to create the concept of “e-Local Governance, that can be defined as “the application of ICTs to transform the business of government and to enable the broad inclusion of citizens in public management, public service delivery and democratic participation at the local level” [14].

5. CASE STUDIES

Contexts and institutional frameworks within African countries are changing rapidly and policy makers and private and public telecommunications service providers have introduced reforms. Reforming countries are reaping the benefits through improved infrastructure, increased applications and better accessibility and affordability of ICTs equipment and services. However, one of the major challenges confronting Africa is to develop the capacity, strategies, and mechanisms necessary to take full advantage of the opportunities offered by ICTs especially at the local level. However, particularly in the African context, the linkages between ICTs and the enhancement of government operations and local good governance is far from being a reality and it is not clear what progress has been made and what are the outcomes in the provision of e-Local Governance in Africa. The four case studies, based upon “field missions” conducted by the author, identify some of the challenges and threats; and what are the good practice strategies and solutions that are emerging in Africa.

5.1 ICTs and Decentralisation, Senegal

Since its independence in 1960, Senegal has committed to a comprehensive process of reform of the State. Crucial to this reform is the policy of “progressive” decentralisation that seeks to empower local government and support local development. The policy of decentralisation is designed in a way that should ensure participation by people and their representatives in the management of public affairs, and this participation should be enhanced as these levels of government become progressively more independent vis-à-vis the central power. Yet there are many obstacles to the effective transfer of these powers. Most local elected officials have no access to legal texts, and they have neither the means nor the capacity to play their roles fully. If we consider the situation of basic equipment and the capacities of managing ICTs, this is even more limited, and in some cases local

governments have not even access to basic infrastructures, such as electricity or telephones. In this context, the African Communities and the Information Society programme, better known simply as ACACIA, an IDRC initiative aiming at “promoting the use of ICTs in African communities, in order to prevent them from being increasingly marginalised from the information society, and also to help them achieve their development objectives”, started in Senegal in October 1996. The strategic justification of the Senegal ACACIA Strategy (SAS) was the decision to make ICTs available to communities, in particular poor communities, and to see how these ICTs could contribute to their development. In fact, despite the then recent advent of the Internet and of high-capacity processing and storage technologies, it was clear to the ACACIA working group that ICTs could enormously support the needs of local governments, as well as the having the potential for increasing participation of citizens in decision-making.

In the light of this situation, the SAS Working Group on Governance launched a research project with the goal to “inform and sensitise local officials, elected representatives and the government about the role and impact of ICTs on decentralisation policy and, at the same time, create the enabling environment for the implementation of an Observatory of Decentralisation and Local Governance”.

The project was executed by SAFEFOD (African Society of Education and Training for Development), a Senegalese non-governmental organisation with a history of experience in local development and governance, with financial support from IDRC/ACACIA, for an amount of 101,276 Canadian dollars for a period of eighteen months (January 1998-May 1999). It was therefore conducted with a reasonable investment in computer hardware and participation of local human resources. The project attained all the goals and expected output of the research, in particular: The Study on ICTs and decentralisation and the “Partnership-Workshop”; The output of the laboratory research: 1) The Web Site - Observatory on Local Governance; 2) Management software applications for local governments: CAURI: Budget Management Software; and CIVIS: Civil Registry Management Software; and a Multilingual Vocal server: French, English, and the six languages spoken in Senegal (Fula, Wolof, Joola, Sereer, Mandinka and Soninke); as well as Demonstration activities and Dissemination of results. As for the technical aspects of the project, it is worth noting the emphasis on the local-based development of contents and application, and the importance given to the “language-divide”. Often, in fact, solutions are just imported from elsewhere and as a result they are not suitable for local requirements or not at all compatible with the technical systems already in place. The issue of language and the capacity to access (accessibility), including the costs, are important factors that, if not solved, can prevent any introduction of ICTs at local level. They should be considered as a “pre” condition for any local governance activity. More importantly, the research was designed in such a way as to provide the local authorities with instruments for managing local government. In fact, the project’s objective was not to amend the decentralization law, but to show how ICTs can help improve its implementation, through partnership between the government and the various players.

In this regard, three major mechanisms for influencing decentralisation policy were addressed: 1) Capacity building: by producing knowledge, through developing applications to meet

different management and organisational needs generated by decentralisation (software for local government management, local budget management, civil registry management, etc.); 2) Broadening the project's influence: by publicising the results and outputs of the project among local officials, and undertaking a study on implementing the observatory on local governance and decentralisation; and 3) Sensitisation and information for local officials, elected representatives and citizens: by providing information about the major laws and regulations governing decentralisation, translation of them into national languages, and posting them online, as well as disseminating the results by demonstrating the products to organisations and potential users. All these activities were conducted during the last phase of the Project and continued afterwards. But, the products and results of the project have not been used with any noticeable impact on public policies at the local level, reflecting in part, the communications gap between research and politics.

In the light of its results, SAFEFOD's Project can definitely be considered "pioneering", and in fact, the requirements that were explored during the research phase are still requirements today. There are currently some attempts to fill the gap, but even now, there is no concrete advancement in the use of ICTs at local government level. A recent survey that has been conducted by the SAFEFOD/OCTIS joint-venture in 43 municipalities in the main region of Dakar, shows that none of them uses a computer to manage the registry system, and there is no specialised software used for the budget management (only a minority of them use Excel software). The results of this survey confirm the same conclusions reached by the IDRC SAS in 1997. It also confirms the difficulty of applying the results of research, especially if innovative due to the context, to produce concrete changes in the governance system. But if we consider that, especially at local level, governance is a highly information-intensive practice, notwithstanding the limitation of government structures, ICTs can be highly beneficial in providing a service to local administration, by making management simpler and more transparent.

The products developed by the project were very simple but extremely useful software to help manage daily procedures. As indicated by the Project manager, "they represent les fondamentaux....(the fundamentals) for any public management. In fact, "How can you govern any administration if you cannot manage correctly and in time the budget, or you don't even know how is the population? Or which is the civil status of citizens?" What has emerged in the analysis of this case, is that in many local governments, there are many other preoccupations that may be considered more important, and so introducing ICTs cannot be considered a "priority among priorities". This is especially true where financial resources are limited and the purchase of ICTs, even if basic, would at the beginning bring more problems than solutions. Who will use it? How will people be trained? What about the costs and time of maintenance and repair if there are problems? How fast do these ICTs become obsolete? What about the cost of replacement? All these questions pose the problem of the return on investment in ICTs that each local administrator has to face when making a decision and, at the same time, facing the consensus of the population that, in many cases, has other more concrete difficulties to solve. This is delaying the process of integration of ICTs into local governments, although, according to the majority of those interviewed, it seems that local governments and their populations are "ready" to introduce and use ICTs. Of

course it is a readiness in terms of not being reluctant, but there is still not readiness in terms of the capacity to manage the ICTs and especially the changes that ICTs bring about. In addition to this, local governments, it must be remembered, are answerable to the central government, and must apply policy measures decided at the national level. In this regard, a clear recommendation that emerged from the case study is the need for a more effective co-ordination of the activities related to decentralisation and local governance, especially when it comes to a significant investment such as the introduction of ICTs. In fact the role of ICTs for local governance being well established, the issue now is how to efficiently manage this process in a cost-effective manner, and how to integrate and co-ordinate the efforts of all different stakeholders involved in the implementation of the decentralisation policy. In this regard, even if it resulted in not being completely successful, the project under analysis demonstrated an interesting approach, where participation and discussion were open and a mechanism of partnership was established to share experiences and ideas and find common solutions. For the successful implementation of the decentralisation policy in Senegal, as well as in many other African countries, this recommendation should definitely be taken into consideration. At present, the Government is revising the National Programme for Good Governance, in order to strengthen the component related to local governance, and its relation with ICTs.

5.2 ICTs and Traditional Governance, Ghana

Ghana has more than 40 ethnic groups and within these groups are a variety of forms and institutions. These include priestly authority, female traditional authority in the forms of Queen Mothers, female chiefs and priestesses, traditional military companies, indigenous health delivery, agriculture and commercial authority. At the apex of these structures of governance is the institution of Chieftaincy, which permeates all through the 40 or so ethnic groups, limited only in power and influence by history, tradition and culture of the various communities where authority is exercised. The pressure due to globalisation and the democratisation process blowing through Africa is now strongly addressing the need to consider the duality of mixed nature of the system of governance. Chiefs are asking for integration of the traditional system into mainstream governance, and at the same time, the role of indigenous institutions, in relation to modern state, is receiving increasing attention. The traditional authorities, as custodians of the land and other natural resources, play a critical role in the economic activities of the people, such as farming, mining, construction, etc. Traditional leaders, as guardians of the history and culture of the people, are thus regarded as one of the crucial echelons of leadership through which the Ghanaian development agenda of poverty reduction and wealth creation could be achieved. A survey in 2001 covering all the regions of Ghana, indicated that there is considerable goodwill towards the institution of chieftaincy among the majority of Ghanaians. In this context, Chiefs are now transforming their role, being often young and well educated people, able to fully grasp the advantages of ICTs and global knowledge.

In this connection, while exploring ways of facilitating its development process through the deployment and exploitation of ICTs within its economy and society, Ghana developed the ICT for Accelerated Development (ICT4AD) Policy, based on an extensive nation-wide consultation with stakeholders from the

public and private sectors, the academic community as well as civil society. In this context, a specific role has been given to the Chieftaincy Institutions that, albeit at the beginning were not fully involved, also thanks to the growing debate around the role of Chiefs as “agents for local development”, have been integrated in the process of development of the ICT4AD Policy and its Strategic implementation. At the same time, the Government of Ghana embarked upon the implementation of a comprehensive Decentralization Policy and local government reform programme in 1988, aiming at establishing a efficient decentralized government machinery as a means to providing strong support for participatory development. The process of decentralization involves a mixture of political devolution enshrined in the constitution, as well as administrative and technical deconcentration of key service delivery institutions which are in part backed by law and in the main reflecting conventional practices. An important aspect of the implementation of the programme is the specific role that is given to civil society organizations, NGOs and private sector, as well as traditional authorities to collaborate in the development of partnership and participate in decentralization efforts. But is it correct to give traditional institutions the same role of civil society organizations?

The debate is open and very controversial. However, the general feeling is that the rather residual functions given to traditional leadership do not afford chiefs the opportunity to play an effective role in modern governance. There is a need to redefine the roles that a traditional leader can play in order to make him/her an integral part of the modern governance paradigm. Taken as a whole, the diversity of the structures of traditional governance institutions and the different phenomena that characterize their development, it is required to establish a framework that can open and facilitate information flow and knowledge sharing. A key component of this framework, that will allow empowerment of chief institutions, maximising their efficiency, in light of its recognition as the embodiment of Ghanaian culture, and as a potential catalyst for development, are ICTs. In fact, the Chief, in his duties as the chief executive of his people, needs ICTs to carry on his day-to-day administrative responsibilities. In their responsibility as development partners, chiefs need management supporting tool and baseline data on their locality and on the available human and material resources. Even when such information exists with the central “modern” government, it is invariably inaccessible to the chief, and yet this is basic to the generation of revenue and attraction of investors. Even more important, is the fact that a vast majority of local communities in Ghana are not “connected”. In many cases there is not even fixed telephone lines and physical transportation is very difficult, especially during the rainy season. It is therefore very complicated to travel and communicate and sending a simple message can take some days to be delivered and is very costly. Introducing ICTs in this context is of great potential advantage both in terms of cost savings and effectiveness. In addition , much of Ghana’s land is vested in the hands of chiefs and traditional councils. The ownership of land and the method of utilization have a far-reaching impact on communities and their people, making traditional governance an important arena for economic development. Chiefs are involved with the exchange of land between persons and groups, thus the Palace of the chief must be able to record such transactions, and there is a clear the need of

computerization of archival records to preserve and disseminate information.

Chieftaincy, as a traditional institution with pre-colonial roots, but which continues to occupy politico-social and cultural spaces in Ghana, as well as in many other African countries The University of Ghana, Legon, Institute of African Studies, rose to the challenge of integrating traditional and modern governance systems, studying and analyzing new ways and tools to support traditional governance and culture for local development. To sustain and support the systematization of these efforts and considering the difficulties and sensitivity that are connected with the topic, the Ford Foundation funded a project on “Chieftaincy, Governance and Development” in the late ‘90s.. This was aimed at researching and documenting the institution of chieftaincy and its role in the general system of governance in the midst of rapid modernization and globalization. With the knowledge gained from this project and other researches and studies, in May 2003, a new Project on “Governance, Culture and Development” funded by the Open Society Initiative of West Africa (OSIWA) has been initiated, to further implement various processes started in connection with the endeavour to modernise Traditional Governance with a more consistent introduction and use of ICTs.

Having realized the importance of records and the manmade and environmental risks of deterioration that they are subjected to, , the Project focused on enhancing the capacities of institutions of traditional governance to enable them participate effectively in mainstream national governance. Critical to this capacity building effort was the introduction of chiefs, and other key players in the chieftaincy institution, to the use of ICTs as tools for modernization and effective governance. The global objective of the Project was therefore to strengthen traditional governance institutions through the introduction of ICTs in their setting and in their operations.

The Project also dealt with the critical interfacing of governance-related cultural issues of concern in Africa, namely: democracy; accountability and transparency; collective responsibility for peace, security and stability; indigenous knowledge; intellectual property rights; cultural preservation as well as the mandatory involvement and active participation of civil society in Africa’s development process. To realize these policy concerns, however, there was obviously a need for greater creative research, training support and advocacy, than is currently offered by a limited portion of the African academic community. In this connection, the uniqueness of the Project is especially in the management of activities, since it has brought together two major developmental partners, academia through the University of Ghana, and the traditional leadership institutions, in an attempt to improve the quality of management and preserve the national cultural heritage through the use of ICTs. Training included capacity building seminars for chiefs and registrars, and other staff on basic computer skills. This is an important task, as the registrars and support staff are mainly documentors, record keepers and information managers in an otherwise oral society. Their ability to provide documentation and effectively manage information serves as a crucial step towards cultural heritage preservation and knowledge building. At the same time, their capacity to produce information, is intended to make the process of traditional governance more accountable and transparent “opening the palace” to the general public.

With regard to partnership between the traditional institutions and the modern state institutions, the introduction and use of ICTs is seen as an effective tool for better cooperation and communication, in order to sustain more effectively local economic development. In general, of course, all this is aimed at solving disputes and preventing conflicts, because in absence of peace it is not possible to have development. In this regard, the role of ICTs is becoming always more and more central. The importance of records of the chieftaincy institutions is evident for a nation that could have avoided a considerable number of conflicts associated with traditional governance disputes, if record had been kept properly for posterity. Digitization and public availability of data will help to create a “memory heritage” from which information can be extracted and used as and when needed. Moreover, it will help to devise and standardise, in a certain way, customary tradition and law. On the other side, the need of equipment and communication facilities is a consequence of the approach., “you cannot learn to drive without driving... and to drive you need a car....”. The technological part of the research thus focused into investigating possibilities of using alternative technologies for establishing a communication system among the traditional governance institutions, using new and affordable technologies. These include the Internet and Satellite communication, especially due the weak accessibility in the remote areas of the country. For this purpose, the possibility to create a WAN (Wide Area Network) using a VSAT connection to offer interoperability among the various institutions across the WAN to the central database was researched. At the same time, a project website (www.chieftaincy.org) was created to enhance networking and promote dialogue among traditional governance institutions, as well as stimulate academic research on governance, culture and development in the country and the rest of Africa and the world. As a consequence of the project, first of all, there is now a growing demand for active involvement of traditional leaders in Central government operations, due to the recognition of the importance of their role in supporting good governance at local level. A clear example is the integration of the Chieftaincy institutions as partners within the “National Governance Programme of the Government”. Several initiatives, in collaboration with developmental partners, such as UNDP, UNECA, the World Bank, have been initiated to assess the situation and the capacity at local level in order to support the Government in better delivering services to the poorest communities, especially in connection with the revitalized process of decentralization. This provides evidence of how the Project, and other related activities undertaken by the University of Ghana, have been initiated a process that could, eventually, bring effective and sustainable development in the country. In this sense, it is also important to underline that the Project aimed at establishing a sustainable process, where institutional and capacity-building is not only a sporadic, project based activity but is in a certain but becomes part of mainstream academic research. It is evident that traditional leaders must become familiar with the modern practice of public sector administration and management while building on the traditions and values that command the trust and respect of members of their communities.

5.3 The Cape Town’s “Smart City” Strategy in South Africa;

For at least the last two decades, there has been significant use of ICT's within government in South Africa for the purposes of improving service delivery to citizens or to enhance back-office. In 1999-2000, there was an acknowledgement that despite, the considerable initiatives in place, there were still many challenges that needed to be addressed if the information systems were to deliver the development priorities of the new state. These included concerns about inter-operability, duplication of efforts, not achieving economies of scale, and security. In addition, the arrangements were not conducive to the creation of seamless access to government services. At the municipal level, the picture is even less clear. Due to the history of fragmentation and the recent creation of the current 284 municipalities, the deployment of ICT's varies substantially. In general, the larger urban metro's have in place systems to manage payments, rates and taxes, registrations, as well as managing their own internal operations. New municipalities and those that exist in marginalized areas are likely to have very few systems in place to assist in local governance. A large scale survey and audit was conducted by a private research agency in 2004 and is currently being repeated (in close association with government). This reveals major ICTs deficits in many marginalized local municipalities including the lack of basic ICTs facilities like a stand-alone computer. In addition, many of these local authorities did not see ICTs as crucial when they were confronting more basic needs challenges like housing, water, sanitation, roads, etc. It has also to be understood against the background that, the new Constitution of the Republic of South Africa (1996) envisaged a complete transformation of the local governance system

In this context, the implementation of any ICTs for development and e-government strategies are likely to have a strong bias towards cities and provincial towns, where the majority of the population reside. But efforts in this regard have yet to deliver major results and therefore it is crucial to focus on approaches and strategies that address the needs of citizens in local government, and in particular in urban and rural and remote areas. Local government should (and can) play a greater role in designing an access strategy as they have the best understanding of the needs of the community, they are responsible for spatial planning, and are playing a major role in overseeing overall socio-economic development. However, there are major capacity constraints in many local authorities and such an approach would require a medium to long term implementation strategy.

In line with the National Strategic Framework, the Provincial Government of the Western Cape (PGWC) has recognized and embraced the important role ICTs plays in poverty reduction and economic growth. In 1998, the Cape IT Initiative (CITI), a not-for-profit networking and cluster development organization, was launched, to bring together people, ideas and capital to grow the Western Cape ICT sector. CITI's goal is to promote Cape Town as a global IT hub and gateway into Africa, thus facilitating the creation of jobs and prosperity through IT. In 2001, the PGWC drafted a White Paper “Preparing the Western Cape for the Knowledge Economy of the 21st Century”, and an e-Government Strategy, “The Cape Online e-Government Programme”.

In order to achieve its goal, the Cape Online Programme has been designed around a number of “core” projects, that address the

internal Government structure, and its capacity to deliver the services. Complementary to these, there are a number of “Online Community Projects”, which are intended to impact various communities of interest, involving specific groups of citizens and organizations, other than the Provincial Government. The Programme is completed by some “External Projects”, which are non-government in nature, and yet impact the online environment for the improvement of business organizations and individuals. To better coordinate the Cape Online Programme implementation, the information technology (IT) and e-government (KEEG) units of the Provincial Government of the Western Cape joined in 2004 to form the Centre for e-Innovation (Ce-I), with the purpose of providing ICTs services to the PGWC, including driving its e-Government strategy. Within this framework, the “Smart City” Strategy of the City of Cape Town, represents an example not only of successful implementation of an ICT-driven reengineering of the City government, but a way to address the twin challenges of poverty alleviation and globalization of the overall provincial government, by identifying the way how ICTs can enable economic and social development and enhance good governance, in the City and in the Province. The multi-award winning Smart City Strategy (it was awarded the “African ICTs achievers e-Government award”, in 2002 and 2003, as well as other), together with a range of complementary city council strategies, represent a possibility to provide an answer to the challenges of the knowledge-economy in South Africa. Through this strategy, the City administration, since 2002, is focusing on providing ICTs skills enhancement opportunities, access to ICTs and business development opportunities. At the same time, it developed a policy discussion document on ICTs and Business Development Services and undertaken a ground-breaking Digital Divide Survey. In fact, recognizing that Digital Divide is having an increasing impact on economic and social development, the City commissioned a Survey to determine whether its communities, businesses and organizations stood in terms of their access to, and use of ICTs. The study underlined that there is a great enthusiasm for ICTs, but a feeling that some communities are being left out of the Information Society.

The vision of the City Council for Cape Town, is “to build a City for all, a City in which no-one is left out”. The Smart City vision, is that of “a Smart City populated by informed people connected to the world and each other by the technology of the information age”. To this end, the Directorates of ICTs, Social Development and Economic Development and Tourism, started implementing a number of projects, from both an externally and internally focused perspectives. The initiatives with the most rewarding results achieved so far, are in the area of internally focused projects. In 2001/2002, the City administration started the rationalisation and standardisation of the IT services within its organisation, as well as enabling internal electronic communications (intranet, emails,); developing the City government Web Site and providing training on ICTs, through a training activity addressed to councillors, to utilise ICTs facilities provided. In 2001, also started the implementation of the Ukuntinga Project – My SAP.com: the largest SAP-ERP implementation and staff training initiative for Local Government in the world, that won the 2004 Computer World Honours Award as the most significant IT project in Government and Non-profit organization.

The Project Ukuntinga (meaning to soar and rise above in a traditional South African language) is about the design and

implementation of an ERP (Enterprise Resource Planning) System that offers a comprehensive solution for managing financial, revenue, human resources, operations and other services (in practical terms its “back office” systems) on a single integrated IT system.

This project has enabled the city to facilitate the merger and transformation of 7 previous autonomous local authorities into a single Unicity. More than 113 legacy systems and 70 interfaces were replaced with a single, functionally rich ERP-SAP system to streamline operations, reduce costs and enhance service delivery. This has been achieved by streamlining and standardizing more than 300 end-to-end business processes on a single integrated transactional system and by training 6,500 staff members to transact on the new system.

The Smart City Strategy, and its ICT-enabled administrative reorganization foundation, the Ukuntinga Project, clearly demonstrate the successful use of ICTs as an enabler of transformation, while merging seven municipal authorities into a single administration with 28,000 payroll members, serving a population of 3.2 million citizens.

This is not purely about the implementation of an IT system, but it is about the relationship between the technology, the city’s business processes and how the organization structures itself around these processes – which the ERP system aims to optimize in support of the strategic objectives of the City. Any change to a single dimension should result in changes to the other two.

The greater reliance on ICT-enabled business processes has fundamentally changed how tasks are performed. One of the benefits of having so many business processes automated is the improved visibility this gives management. A breakdown in a process or a backlog in a service can now be managed and remedial actions implemented based on objective indicators. The city can now monitor its service offerings centrally, but has the freedom to deliver them through devolved structures. Remedial action can be taken pro-actively, based on objective measures and no longer will the number of people who complain or the length of the queue be the only insight into the demands being placed on a service. At the same time, the uniqueness of many aspects of the system can be identified in the support to the City’s social and developmental objectives and, in the fact that, although the project was implemented in record time it was conceived through a deliberate process spanning many years and with the involvement of all the key stakeholders.

However, as in many cases in Africa, the City of Cape Town operates as two distinct societies – one wealthy and developed, and the other poor and underdeveloped. The key challenge faced by the city is the notion of “inclusiveness”. The poor and disadvantaged must be catered for no less than anybody else. This has generally not been the case. By the same token, the city recognises that an approach that seeks to cater for the special needs of the poor and vulnerable in a manner which impacts unreasonably on the interests of any other segment of the city is also not supported.

A serious concern surrounding the introduction of such an ERP system, is that it is a “first world” solution and only benefits the wealthy, developed communities in the city. This can result in exacerbating the “digital divide” across the city, and in the process leave poor communities further behind. As a city, Cape Town needs to ensure that the poorest communities are able to

keep pace with the latest technology developments. Hence, the approach followed has been focused on solutions that provide equivalent benefits to all communities, addressing the needs of the broader society, rather than focused on the needs of any specific segment. A recent example of this was the translation of the municipal accounts from English into a local, traditional language, isiXhosa. But the Ukuntinga Project has to be considered as part of an holistic strategy that, in addition to reorganising the back office for better service delivery, seeks to address the “digital divide” challenge by, for example, providing free access to computers and the Internet to the people of Cape Town at municipal libraries (the city’s Smart Cape Access Project, www.smartcape.org.za that won the Bill and Melinda Gates Access to Learning Award 2003); or providing training on ICTs to disadvantaged people (the Kulisa Project). A further example of this challenge is the ability to get small, micro and medium enterprises to embrace information technology in order for them to do business more effectively with the city (Library Business Corners; and Digital Business Centres).

Although many people referred to the project as an ERP implementation project, it was in essence a transformation project, driven by the technology implementation. The technology forced the execution and implementation of the new processes and organisational design, and its administrators to make and implement decisions that may have been postponed, were it not for the aggressive implementation plan. It is therefore important to underline the need to have a strong and capable leadership capacity, as well as the involvement of the necessary expertise. In this regard, Ukuntinga Project is instead a showcase in terms of the development of historically disadvantaged individuals. Through the innovative partnering and deal structure between the City of Cape Town and the international consulting and technological private company contracted, the project provided the opportunity for more than a hundred historically disadvantaged individuals and small and medium sized companies to build ERP and SAP capacity. These individuals derived methodologies and knowledge, from the private companies’ expertise, by working side by side with the consultants during the project .

5.4 The District Administrative Network Programme in Uganda

ICTs in local governance in Uganda have been identified as a major tool for achieving socio-economic development. In order for government to implement the long term national development programmes, timely and relevant information must be available at all levels of implementation. The National ICT Policy Framework (2003) is intended to stimulate more participation in the socio-economic-political and other developmental activities, so as to lead to improved standards of living for the majority of Ugandans and enhance sustainable national development. In March 2005, the ICT/e-Government Inter-Agency National Planning Team was established, under the co-ordination of the National Planning Authority. The main goal of the Team is to discuss ways and means of ensuring that ICT/e-Government services are undertaken by the Government of Uganda as one of the core priority sub-sectors in the execution of the proposed “National Vision 2035: Towards a Modern Industrialised and Knowledge Based Society”. At the same time, decentralisation in Uganda has been implemented for over 12 years now, guided by the Local

Government Statute of 1993, the 1995 Constitution and the 1997 Local Government Act. This is based on the conviction that decentralisation comprehensively facilitates the realisation of developmental and political objectives for Uganda through democratisation, equitable distribution of resources among and within districts and improvements in the public sector performance.

The system of Local Government in Uganda is based on the District as a Unit under which there are lower Local Governments and Administrative Unit Councils (at present there are 56 Districts with over 900 sub-counties). Districts, counties, sub-counties and parishes have been empowered by the Uganda “Local Government Act 1997” to be self-governing. The overall management of a local council now requires that the council make considerable investment in human resources and infrastructure to manage these increased responsibilities. A lot of financial resources (from both donors and the Government) is financing various projects at local council level in various districts. The concern in all cases has been that financial resources must be used for the purpose for which they were intended. To discuss these issues, a Roundtable Workshop with the theme “ICT for Rural Development” was held in March 2001, in Jinja, with participants from rural and upcountry institutions in public and private sectors. Following the Round Table, the District Administrative Network Programme (DistrictNet) was approved for implementation at the district headquarters and lower local governments and councils, in four pilot districts of Kayunga, Lira, Mbale and Mbarara., selected from each of the four ‘regions’ of Uganda. DistrictNet is managed by the Districts themselves, with the support of the Ministry of Local Government, as pioneers for the benefit of a possible national approach in due course. External funding was secured by the International Institute for Communication Development (IICD) in collaboration with the Department for International Development (DFID) of the United Kingdom to finance the first/pilot year’s investments and some of the operational costs. A contribution from the Ministry of Local Government and the beneficiary Districts was also required, bringing the total cost of the project to 921.839 Ugandan Shillings. The project aim is to improve the performance of local Governments by establishing functional data/information management and public communication systems for effective and efficient decentralised service delivery. During the first year, the project was planned to be implemented in the four pilot district headquarters and 11 sub-counties. Afterwards, it was expected that the benefits of the project would emerge, and useful lessons learned. The project also provided for a seminar to share experience and lessons learned, with other districts, but it was not within the scope of the project to roll out to all districts and sub-counties. However, 11 more sub-counties were planned to be covered during the second year, and a further 21 sub-counties in the third year. Moreover, considering that many government programmes and projects are already focusing on building the capacity of Local Governments to manage their affairs more efficiently and more transparently, the project proposed to set up a Steering Committee to, among other things, liaise with other Districts and other Ministries to ensure rollout of the project to the remaining sub-counties and districts. In the pilot phase, the project was planned to introduce: data and voice communication links between districts and lower local governments; and electronic data processing in financial management, data communication, storage and analysis.

So far, the four pilot districts where the project has been implemented have made savings in administrative expenses thus freeing these funds to be used for other more pressing economic activities geared towards economic development initiatives. This is due also because, in addition to improvement in the communication channels, users have now easier access to useful information for planning purposes. In social terms, the communities in the districts covered by the project have been sensitised to the usefulness of utilising ICTs and how these technologies can bolster development. One of the main results is that there is now an increasing demand for accurate and timely information from technical staff by the politicians to support their decision-making functions. Thus, higher levels of ICTs awareness are now helping development and this success has been duly noted: Members of Parliament have in fact promised to ensure the resources to roll out District Net to the rest of the country.

The project has now achieved a strong appeal among other districts. Reinforcing ICTs infrastructure and capacities is in fact viewed as a key component in the development of local governance systems. Once the transformation of the districts' administrative structures and the general operating systems have been completed, the districts are in a position to communicate fully with sub-counties and the central government through increased use of computers for documentation, storage, transfer of information and file sharing. This allows for a reduction in the costs of communication (e.g. transport and production of documentation) and an increase of public information and service delivery, as well as easier and more accurate data collection through the use of standardised pre-designed forms. All this will give constituents (citizens and private companies) the possibility of better interaction with the local government, thus enhancing their capacity to produce socio-economic development at the local level.

Despite the success of the project in its pilot phase and the interest in replicating similar initiatives in other districts and some sub-countries not covered., however it is critical that the pilot project demonstrates success by the completion of the implementation. A number of obstacles have been encountered in the implementation of activities, delaying its completion. One of the main problems is related to delays in the procurement process (Government procurement systems) that also resulted in delays of contractual arrangements and in the supply of equipment by some providers. The financial resources proved to be not adequate enough to purchase technical equipment not previously budgeted for (i.e. technical support equipment for maintenance purposes), and there were limited options for connectivity solutions especially in sub-counties. Thus, Internet connectivity is still a problem in all the districts although it has been tested in Kayunga, Lira and Mbale. In the specific case of the Lira district, security problems threatened the effective implementation of activities. In particular, despite the need to link up with the sub- counties, technology compatibility has still not been achieved, and so effective connection with sub-counties is not yet a reality.

By the start of 2005, the project had entered into a mainstreaming phase, and now acquired private sector partners in addition to local and central government, IICD and DFID. As a result of the overall implementation of the project, it can be said that today, there is a fast growing realisation in Uganda that e-Governance can bring individuals into close contact with decision makers and officials in the government, especially at the local level. But e-

Local Governance can only be implemented effectively if the right human, technological and financial resources, are available and the citizens are ICT literate and sensitised. Thus, the government is currently in the process of acquiring more funding to extend the project to other districts. This is based on the lesson that is now learned by many projects all over the world "think big, start small, scale fast"! The pilot activities are in fact now under analysis in order to be reproduced in other projects and implemented in other districts through a national programme. However, the growth and significance of the project will depend on the effort by the Ministry of Local Government, working together with other ministries and districts to integrate all operations and in particular the database to avoid duplication. It is therefore important that the Ministry of Local Government will maintain control of the design and choice of technologies so as to ensure standardisation and interoperability. It is also pivotal to reinforce the dissemination activities through advocacy campaigns and specific national and local workshops. In particular, the importance of data collection and information management should be raised, especially at the lower level of government.

A hopeful result of the process in action to "scale up", District Net has the possibility to create a "multiplier effect", encouraging local governments (both districts and sub-counties), to invest their own resources in ICTs equipment and training, as well as developing planning and assessment activities at the lower local level. To have a greater impact, it is also envisaged that when designing the projects, districts and the national government will keep following the participatory model that is at the heart of "District Net", and including all stakeholders and beneficiaries. The current approach of "networking-communities" developed for example in the "I-Network Uganda", as well as the national Inter Agency Team, is promising, but there is still the need to undertake research in the appropriate technology (software) that is free or affordable, especially for marginalised groups and areas, as well as how ICTs can improve health service delivery especially among the poor in the rural areas. These could be some additional components of a further, broader "District Net" Programme in Uganda.

6. CONCLUSIONS

The main lessons learned from case studies and empirical analysis show that, first of all, there is no single way of introducing ICTs. The process is dynamic and consists of several stages, especially in Africa: raising awareness about the potential of ICTs for community development; encouraging basic use of ICTs; providing specific products and content to meet local demands (e.g., materials in national languages and products tailored to the needs of specific sectors of the population).

Participation is a crucial problem in the process of introducing and promoting the use of ICTs for community development. Appropriate mechanisms should have been initiated within the communities, but finding ways to involve large segments of the population still constitutes a real problem, even when people are aware of the potential usefulness of ICTs.

Due to installation costs and the recurrent expenses involved in the use of ICTs (i.e., Internet and email), alternative technologies (e.g., satellites with wireless technology and multimedia tools) should be considered in introducing ICTs, in order to better adapt to the infrastructure available and so improve community access.

Moreover, country specificity and the institutional context to ICTs project implementation, must be taken into appropriate consideration. In this connection, political will, community leadership and ownership are key enabler factors, as well as accurate strategic planning, effective monitoring and critical evaluation are indispensable to identify factors inhibiting impact and ensure sustainability. A first aspect to consider, concerns the fact that local languages and illiteracy constitute a barrier to access of information. In this connection, it is rightly believed that, unless ICTs are made available in local languages, the ongoing ICTs revolution will remain incomplete and its benefits are likely to reach only a small section of society who have access to the “linguafranca” of the Internet (English).

The scenario is, however, undergoing a fast change. For instance efforts are being continuously made in other countries (e.g. in India), some with remarkable successes, in making ICTs available in local languages, often called regional languages, thus taking the benefits of ICTs to the common man and woman. In Africa this is an even more complicated issue, considering the number of local languages and dialects that are spoken, and the dual system of governance still existing in many countries. Local content is invariably available in the form of indigenous and traditional knowledge that has been inherited by the community over centuries. One aspect of this is that ICTs could provide the critical tools for launching traditional governance into the information age, bringing about all the advantages that they offer for development and good governance, using the traditional leaders and chieftaincy institutions as promoter and catalyst for innovation and local community development.

However, the gap between connectivity and technology capacity on the one hand and content on the other grows ever more vast. More so, in African countries, where the mastery of technology seems to be an end in itself, almost wholly divorced from the need to solve the many problems of deprived millions.

The availability of the appropriate skills base is a crucial determinant of the growth of ICTs supply activities, and these contribute to the more general human resource development. At the same time, the skills base must be understood as an important risk factor in appraising communication network infrastructure expansion and ICTs applications projects. Without available skills to operate and maintain the physical infrastructure, as well as develop and maintain software, users or potential users will naturally be unable to take advantage of the infrastructure, which as a consequence, will not be used to its full potential. Another potential risk factor related to training is the “brain-drain” risk. Highly trained individuals must be given posts which fulfil their expectations and ambitions.. Thus it is important that capacity building will not only focus on the training of individuals, but on reinforcing the capacity of organisations (private and public) especially the institutions at local government level to make good use of them.

From all these considerations, it appears evident that the challenge remains as to how the use of ICTs in local government can be beneficial to all the stakeholders, taking into consideration real factors such as the digital divide (both international and domestic) and the prohibitive cost of traditional technologies; it remains a challenge to full citizens’ participation in order to create an “e-inclusive” society. What is difficult is not introducing technologies, but how people can use the technologies best !

7. REFERENCES

- [1] 4th Global Forum on Re-inventing Government - Citizens, Businesses and Governments: Dialogue and partnerships for Development and Democracy, Marrakech, Morocco, 10-13 December 2002, UN, New York 2002. www.unpan.org;
- [2] See, Finger Matthias, Conceptualizing e-Governance, European Review of Political Technologies, March 2005;
- [3] Cheema Shabbir and Maguire Linda, “Democracy, Governance and Development: A Conceptual Framework”; 4th Global Forum on Re-inventing Government, Marrakech, Morocco, 10-13 December 2002, UN, New York 2002.
- [4] Kauzya John-Mary, “Local Governance Capacity Building for Full Range Participation: Concepts, Frameworks, and Experiences in African Countries”; 4th Global Forum on Re-inventing Government, Marrakech, Morocco, 10-13 December 2002, UN, New York 2002
- [5] Cheema Shabbir and Maguire Linda, see above
- [6] Saxena, K.B.C. “Towards Excellence in e-Governance” International Journal of Public Sector management, Volume 18 N.6, 2005, Emerald Group Publishing.
- [7] This definition and concept of e-Governance has been developed by Gianluca Misuraca for the African Training and Research Centre in Administration for Development (CAFRAD) within the framework of the e-Africa Initiative for Good Governance: Building e-Governance capacity in Africa, Project Proposal, 2002; and further refined within the framework of the 2005 Executive Master in e-Governance at EPFL-Ecole Polytechnique Fédérale de Lausanne (<http://egov.epfl.ch>). See also, Misuraca Gianluca, “Research on ICTs for Local Governance in Africa”, IDRC, UNECA (to be published). See also, Misuraca G., “e-Africa Initiative for Good Governance: Building e-Governance Capacity in Africa”, Encyclopedia of Developing Regional Communities with Information and Communication Technology, IDEA Group, 2005. It is to be considered a starting point of a longer journey towards a fully developed definition of e-Governance.
- [8] UNESCO, www.portalunesco.org
- [9] Misuraca Gianluca, “From e-Government to e-Governance: the e-Africa Initiative for Good Governance”, paper presented to the International Conference on e-Government, SITEXPO 2004, Casablanca, Morocco, February 2004.
- [10] e-Government Policy Network of the Privy Council Office (PCO) “Transforming Government and Governance for the 21st Century”, www.publiservice.pco-bcp.ga/egov-cybergouv, 2003
- [11] For a recent discussion see above K.B.C. Saxena.
- [12] Pablo D. Zelinna, and Pan L. Shan, “A Multi-Disciplinary Analysis of e-Governance: Where Do We Start ?”, La Salle University, Philippines, & National University, Singapore. 2003
- [13] UN Road map to good governance and democracy www.unpan.org
- [14] LOG-IN Africa, Project Proposal to IDRC, Misuraca Gianluca, on behalf of CAFRAD, October 2005.

Automated Classification of Congressional Legislation

Stephen Purpura

John F. Kennedy School of Government
Harvard University
+1-617-314-2027

stephen_purpura@ksg07.harvard.edu

Dustin Hillard

Electrical Engineering
University of Washington
+1-206-789-1029

hillard@ee.washington.edu

ABSTRACT

For social science researchers, content analysis and classification of United States Congressional legislative activities have been time consuming and costly. The Library of Congress THOMAS system provides detailed information about bills and laws, but its classification system, the Legislative Indexing Vocabulary (LIV), is geared toward information retrieval instead of the pattern or historical trend recognition that social scientists value. The same event (a bill) may be coded with many subjects at the same time, with little indication of its primary emphasis. In addition, because the LIV system has not been applied to other activities, it cannot be used to compare (for example) legislative issue attention to executive, media, or public issue attention.

This paper presents the Congressional Bills Project's (www.congressionalbills.org) automated classification system. This system applies a topic spotting classification algorithm to the task of coding legislative activities into one of 226 subtopic areas. The algorithm uses a traditional bag-of-words document representation, an extensive set of human coded examples, and an exhaustive topic coding system developed for use by the Congressional Bills Project and the Policy Agendas Project (www.policyagendas.org). Experimental results demonstrate that the automated system is about as effective as human assessors, but with significant time and cost savings. The paper concludes by discussing challenges to moving the system into operational use.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Clustering, Information Filtering, Retrieval Models

General Terms

Algorithms, Performance, Experimentation

Keywords

U.S. Congress, legislative activities, text analysis, SVMs, support vector machines, institutions.

1. INTRODUCTION

The Congressional Bills Project received NSF funding in 2000 (SES 0080061) to assemble a dataset¹ of all federal public bills introduced since 1947. The project's data set contains 390,000 records that include details about each bill's substance, progress and sponsors. Each bill is also assigned a single topic code drawn from the 226 subtopics of the Policy Agendas Project². The resulting database is of high quality and used by researchers, instructors, students and citizens to study relative policy attention across time and venues. Researchers on other project teams are also classifying other government, media and public activities according to the same system, expanding the scope of comparison. A subset of published research, including articles and books, that consume the data may be found at the Policy Agendas web site³.

At this time, a common classification scheme from the Policy Agendas Project makes possible comparisons of all Congressional bill activity with all Congressional hearings activity, Presidential State of the Union addresses, New York Times stories (sample), Solicitor General Briefs, and Gallup's Most Important Problem poll indices, among others for the period 1947-present. To date, these classification projects have depended on the efforts of trained human coders. However, the time and cost involved in expanding to new datasets and continually updating existing systems are substantial. A high quality, automated approach, especially one that allows lessons learned in one venue to be applied to another, would greatly speed the availability of the data to researchers.

Unfortunately, published attempts detailing the development of automated sorting and classification tools for projects of this scale and complexity are few. Recent research from Benoit, Laver, and Garry [7] has examined automated classification of issue appeals in party platforms using a word scoring technique. In addition, Shulman and others [6][12] have examined regulatory comment email duplicate detection using Kullback-Leibler (KL) distance and clustering techniques. Although Shulman's work is closer to our approach, we will instead propose a general purpose method borrowed from research in newswire topic spotting in computational linguistics.

¹ See www.congressionalbills.org

² See www.policyagendas.org and the codebook at:
<http://www.policyagendas.org/codebooks/topicindex.html>

³ See <http://www.policyagendas.org/publications/index.html>

On first appearance, legislative bills have similar document characteristics to newswire data. Topic spotting in legislative bills has similar goals to topic spotting in newswire data because both involve scanning a text segment for the predominance of a theme. Numerous techniques for topic classification have been well documented. In this work, support vector machines (SVMs) are chosen due to their strong performance on a wide variety of tasks.

SVMs are a natural fit for topic classification because they deal well with sparse data and large dimensionality. But legislative text has different language patterns and characteristics from the typical news stories or broadcasts usually classified in newswire topic spotting. Unlike news stories or broadcasts, legislative text uses a standard template and the language may be very similar for specific types of bills. We propose the commonalities will overwhelm the difficulties and make the task of topic spotting in legislation quite successful.

The remainder of this paper documents our approach to building a prototype of a SVM system to classify the legislative text of the U.S. Congress using the Policy Agendas coding scheme and human coded samples. The approach was tested on roughly 108,000 of the 390,000 records in the Congressional Bills Project databases, as this was the largest sample available at the time of analysis. The approach to classifier design is developed in Section 2. The evaluation methodology is presented in Section 3. Experimental results are detailed in Section 4, and the main conclusions of this work are summarized in Section 5.

2. ALGORITHM OVERVIEW

Our goal is a software system that assists the Congressional Bills Project in classifying bills from the U.S. Congress according to the Policy Agendas coding scheme. Based on training examples (known as ‘the truth’) from expert coders, the system should scan each bill and determine which of 226 subtopic codes best fits each bill. The section below describes an algorithm that accomplishes the objective.

2.1 Support Vector Machines

SVMs were introduced in [14] and the technique attempts to find the best possible surface to separate positive and negative training samples. The best possible surface produces the greatest possible margin among the boundary points.

SVMs were developed for topic classification in [4]. Joachims motivates the use of SVMs using the characteristics of the topic classification problem: a high dimensional input space (the words), few irrelevant features, sparse document representation, and the knowledge that most text categorization problems are linearly separable. All of these factors are conducive to using SVMs because SVMs can train well under these conditions. That work performs feature selection with an information gain criterion and weights word features with a type of inverse document frequency. Various polynomial and RBF kernels are investigated, but most perform at a comparable level to (and sometimes worse than) the simple linear kernel. A software package for training and evaluating SVMs is available and described by [5]. That package is used for these experiments.

2.2 Word Feature Processing

Text input to topic classification systems is usually preprocessed and then word features are given weights depending on importance measures. Most text classification work begins with word stemming to remove variable word endings and reduce words to a canonical form so that different word forms are all mapped to the same token (which is assumed to have essentially equal meaning for all forms). Word features usually consist of stemmed word counts, adjusted by some weighting. Inverse document frequency is commonly used, and has some justification in [8]. More complex measures of word importance have shown to provide additional gains though. A weighted inverse document frequency is an extension of inverse document frequency to incorporate term frequency over texts, rather than just term presence [11]. Term selection can also help improve results and many past approaches have found information gain to be a good criterion ([13] and [10]).

During word feature processing, we remove non-word tokens, map text to lower case, and then apply the Porter Stemming Algorithm described in [9]⁴. The text is then distilled into features. Features such as inverse document frequency have been generally effective but more detailed forms of word weighting have shown improvements. This work adopts a weighting related to mutual information. Each word is given a feature value w_i as shown in equation 1.

$$w_i = \log\left(\frac{p(w,t)}{p(w)p(t)}\right) = \log\left(\frac{p(w|t)p(t)}{p(w)p(t)}\right) \quad (1)$$

In this equation, the top term, $p(w|t)$, is the probability of a word in a particular bill (the number of occurrences in this bill, divided by the number of total words in the bill). The denominator term $p(w)$ is the probability of a word across all bills (the number of occurrences of this word in all bills, divided by the total number of words in all bills). This also reduces to an intuitive form as in equation 2 where it can be thought of as a ratio of word frequency given a bill, divided by the overall frequency in all available bills.

$$w_i = \log\left(\frac{p(w|t)}{p(w)}\right) \quad (2)$$

Finally, only words with $w_i > 0$ are placed in the term by conversation matrix (this is all terms with a ratio greater than 1, or in other words those that occur more frequently than the corpus average).

2.3 Hierarchical Approach

Our approach is unique because our problem demands innovation on the typical use of SVMs. We have chosen a two-phase hierarchical approach to SVM training which mimics the method employed by human coders. Human coders first classify a bill as falling under one of 20 major topic codes (see Table 1) and then further classify it as falling under one of 226 subtopics. For example, a bill proposing to reform the health care insurance system is assigned to fall under subtopic 301, where the 3 indicates health, and the 01 indicates health insurance reform.

⁴ Note that this step reduces performance in international environments. See discussions of stemming.

Table 1: Major Topic Codes

1 = Macroeconomics
2 = Civil Rights, Minority Issues, and Civil Liberties
3 = Health
4 = Agriculture
5 = Labor, Employment, and Immigration
6 = Education
7 = Environment
8 = Energy
10 = Transportation
12 = Law, Crime, and Family Issues
13 = Social Welfare
14 = Community Development and Housing Issues
15 = Banking, Finance, and Domestic Commerce
16 = Defense
17 = Space, Science, Technology, and Communications
18 = Foreign Trade
19 = International Affairs and Foreign Aid
20 = Government Operations
21 = Public Lands and Water Management
99 = Other

The advantages of the two phase approach were many, but two reasons stand out. First, training SVMs on 226 subtopic codes across large numbers of bills is computationally expensive. Using this hierarchical approach greatly reduces the computational expense of the sorting. The hierarchical approach can be implemented on a common laptop computer with a complete sorting of the full data set in much less than a day of processing. Second, human coders are more likely to disagree on subtopic coding than they are on major topic coding. Thus, correctly predicting the major topic of a bill has more value to the coding team than completely missing the mark.

The hierarchical approach's two-phase system begins with a first pass which trains a set of SVMs to assign one of 20 major topics to each bill. The second pass iterates once for each major topic code and trains SVMs to assign subtopics within a major class. For example, we take all bills that were first assigned the major topic of health (3) and then train a collection of SVMs on the health subtopics (300-398). Since there are 20 subtopics of the health major topic, this results in an additional 20 sets of SVMs being trained for the health subtopics.

Once the SVMs have been trained, the final step is subtopic selection. In this step, we assess the predictions from the hierarchical evaluation to make our best guess prediction for a bill. For each bill, we apply the subtopic SVM classifiers from each of the top 3 predicted major topic areas (in order to obtain a list of many alternatives). This gives us subtopic classification for

each of the top 3 most likely major categories. The system can then output an ordered list of the most likely categories for the research team.

3. EVALUATION METHODOLOGY

Evaluation of success is straightforward because high quality information which describes “the ground truth” is available. This section describes the data sets used in our experiments and our methodology for assessing performance against human labelers.

3.1 Data Sets

This research was conducted using the Congressional Bills Project’s public data set⁵. At the time (April 2004), ‘only’ 108,000 records were available for analysis. All statistics are generated from the 108,000 record set.

For the purposes of testing, the 108,000 records were divided into two groups and processed using the “train on 50%, test on 50%” methodology. We report results for the entire set using cross validation, which means we run the system twice (the second run swaps the train and test examples), allowing us to test on all available bills. To select the groups, random sampling without replacement was applied across all of the bills. The experiment was repeated many times, and the statistics were comparable. We report the last run.

3.2 Evaluation Metrics

We use metrics common in topic spotting and clustering analysis work in our evaluation of performance. The usefulness of our system was measured by its ability to predict the truth for every record. For analysis convenience, we also summarize consistency with the truth by major topic and subtopic classifications. Finally, we report Cohen’s Kappa and AC1 to assess inter-coder agreement with the human team, as described in [3] and [12].

Cohen’s Kappa statistic is a standard metric used to assess inter-coder reliability between two sets of results. Usually, the technique is used to assess results between two human coders, but the computational linguistic field uses the metric as a standard mechanism to assess agreement between a human and machine coder.

Cohen’s Kappa statistic is defined as:

$$\kappa = \frac{p(A) - p(E)}{1 - p(E)} \quad (3)$$

In the equation, $p(A)$ is the probability of the observed agreement between the two assessments:

$$p(A) = \frac{1}{N} \sum_{n=1}^N I(Human_n == Computer_n) \quad (4)$$

Where N is the number of examples, and $I()$ is an indicator function that is equal to one when the two annotations (human

⁵ Data is available from www.congressionalbillsproject.org

and computer) agree on a particular example. $P(E)$ is the probability of the agreement expected by chance:

$$p(E) = \frac{1}{N^2} \sum_{c=1}^C (HumanTotal_c \times ComputerTotal_c) \quad (5)$$

Where N is again the total number of examples and the argument of the sum is a multiplication of the marginal totals for each category. For example, for category 3, health, the argument would be the total number of bills a human coder marked as category 3, times the total number of bills the computer system marked as category 3. This multiplication is computed for each category, summed, and then normalized by N^2 .

For reasons of bias documented by [3], computational linguists also use another standard metric named the AC1 statistic to assess inter-coder reliability. The AC1 statistic corrects for the bias of Cohen's Kappa by calculating the agreement by chance in a different manner. It has similar form:

$$AC1 = \frac{p(A) - p(E)}{1 - p(E)} \quad (6)$$

But the $p(E)$ component is calculated differently:

$$p(E) = \frac{1}{C-1} \sum_{c=1}^C (\pi_c (1 - \pi_c)) \quad (7)$$

Where C is the number of categories, and π_c is the approximate chance that a bill is classified as category c .

$$\pi_c = \frac{(HumanTotal_c + ComputerTotal_c)/2}{N} \quad (8)$$

In this paper, we report both Cohen's Kappa and AC1 because the two statistics provide consistency with topic spotting research and most other research in the field. For coding problems of this level of complexity, a Cohen's Kappa or AC1 statistic of 0.70 or higher is considered to be very good agreement between coders.

4. EXPERIMENTAL RESULTS

The Congressional Bills Project assessed the system by its ability to reliably predict the major topic and subtopic about as well as a human. These results are reported in Tables 3 through 6, and they express that the system is about as accurate as a trained human coder at identifying the major topic of a bill, and sometimes as accurate at identifying the subtopic of a bill, with some exceptions.

The results in Table 2 illustrate that the system automatically determines the correct major category for over 80% of the bills. The single worst category is Category 99, which makes sense because this is an 'Other' category only used for bills that could not reasonably be assigned to any other category. Performance on other categories varies, but is mostly above 80% correct. The single best category was Category 18, 'Foreign Trade' at almost 90%. Excluding the 'Other' category, the most difficult category

Table 2: Major Category Precision; Number of Bills Predicted Correctly by Major Category, including totals.

Category	Correct	Possible	Percent
Macroeconomics (1)	4148	5481	75.68
Civil Rights ... (2)	1682	2397	70.17
Health (3)	7246	8200	88.37
Agriculture (4)	3137	3703	84.72
Labor ... (5)	5232	7323	71.45
Education (6)	3131	3613	86.66
Environment (7)	4108	4871	84.34
Energy (8)	4128	4660	88.58
Transportation (10)	4518	5378	84.01
Law, Crime ... (12)	5417	6491	83.45
Social Welfare (13)	5249	6080	86.33
Community ... (14)	1851	2447	75.64
Banking ... (15)	5261	6876	76.51
Defense (16)	6255	7440	84.07
Space, Science (17)	1500	1845	81.30
Foreign Trade (18)	4127	4647	88.81
International (19)	1613	2372	68.00
Government Op (20)	13416	15607	85.96
Public Lands ... (21)	6830	7894	86.52
Other (99)	145	943	15.38
Total	88994	108268	82.20

Table 3: Subcategory Precision; Number of Bills Predicted Correctly for Subtopic Categories (totals only).

Subtopic	Correct	Possible	Percent
Total	76800	108143	71.02

was Category 19, 'International Affairs and Foreign Aid' at only 68% correct.

Table 3 presents the overall statistics for categorization at the subtopic category level. The number of possible bills is slightly lower (only by 0.1%) because our hierarchical approach only hypothesizes minor categories within the top three major categories for each bill. This provides for significant computational savings, while missing only a negligible number of bills. The overall percentage of correct bills is 71% and is lower than for the major categories, but this task is significantly more complex with over 200 possible categories instead of 20 for the major category case.

Tables 4 and 5 present the 15 best and worst individual minor category results. The single best category is 1807 'Tariff and Import Restrictions, Import Regulation.'

Table 4: Subcategory Precision; Number of Bills Predicted Correctly for Subtopic Categories (best 15 subtopic categories).

Category	Correct	Possible	Percent
Tariff and Export Restrictions (1807)	2754	2974	92.60
Federal Holidays (2030)	322	351	91.74
Relief Claims Against the U.S. Government (2015)	3071	3378	90.91
Airports, Airlines, Air Traffic Control, and Safety (1003)	1022	1155	88.48
Food Stamps, Food Assistance, and Nutrition Monitoring Programs (1301)	520	591	87.99
Regulation of Political Campaigns, Political Advertising, PAC Regulation, Voter Registration, Government Ethics (2012)	1257	1447	86.87
Worker Safety and Protection, Occupational and Safety Health Administration (OSHA) (501)	470	542	86.72
Government Subsidies to Farmers and Ranchers, Agricultural Disaster Insurance (402)	1379	1594	86.51
Highway Construction, Maintenance and Safety (1002)	623	721	86.41
Tobacco Abuse, Treatment, and Education (341)	258	299	86.29
Broadcast Industry Regulation (TV, Cable, and Radio) (1707)	538	624	86.22
Natural Gas and Oil (Including offshore Oil and Gas) (803)	1532	1783	85.92
Recycling (707)	176	205	85.85
Postal Service Issues (including Mail Fraud) (2003)	806	942	85.56
Native American Affairs (2102)	854	1009	84.64
Higher Education (601)	1397	1653	84.51

Many of the minor categories that had a large number of examples had better performance in the end, probably because the SVM was better able to learn the category characteristics when more examples were available. The 15 worst categories are primarily those categories with very few examples, and often were again those categories that were ‘Other’ categories within a major topic (those ending in 99).

Table 5: Subcategory Precision; Number of Bills Predicted Correctly for Subtopic Categories (worst 15 subtopic categories)

Category	Correct	Possible	Percent
Unemployment Rate (103)	0	17	0.00
Social Welfare, Other (1399)	0	39	0.00
Banking, Finance, and Domestic Commerce, Other (1598)	0	6	0.00
Foreign Trade, Other (1899)	0	14	0.00
Anti-Government Activities (209)	0	17	0.00
Public Lands and Water Management, Other (2199)	0	6	0.00
Drugs and Alcohol or Substance Abuse Treatment (344)	0	42	0.00
Education Research and Development (698)	0	15	0.00
International Affairs and Foreign Aid, Other (1999)	1	23	4.35
Military Nuclear and Hazardous Waste Disposal, Military Environmental Compliance (1614)	2	41	4.88
Energy, Other (899)	1	17	5.88
Other, Other (9999)	65	863	7.53
Transportation, Other (1099)	2	26	7.69
Labor, Employment, and Immigration, Other (599)	3	29	10.34
Civil Rights, Minority Issues, and Civil Liberties, Other (299)	2	19	10.53

4.1 Systems-to-Human Inter-coder Agreement

The second set of calculations assessed inter-coder reliability, as calculated using Cohen's Kappa and AC1. We use a single coder to express the performance of the entire Congressional Bills team and note that in future research we will integrate the system as a coder within the team for testing. The calculations are summarized in Table 6, and demonstrate, using either Cohen's Kappa or AC1 as metrics, the system performs about as well as humans would be expected to perform.

TABLE 6: Cohen's Kappa and AC1, humans versus system

	p(A)	p(E)	Statistic
κ for all major topics	0.822	0.069	0.809
κ for all subtopics	0.710	0.013	0.706
AC1 for all major topics	0.822	0.049	0.813
AC1 for all subtopics	0.710	0.004	0.709

5. CONCLUSION AND NEXT STEPS

Researchers are now classifying government, media and public activities according to common coding systems to expand the scope of comparison across government institutions. The Congressional Bills Project and the Policy Agendas Project are just two examples. Their experience makes clear that the shift from paper documents to electronic documents should make their job easier, but without new tools and methods, progress will be slow and expensive.

This research focused on the process of sorting United States Congressional bills using an established classification system. Extensive work by the Congressional Bills team set the benchmark for measuring an automated system. And the techniques in this paper demonstrate that support vector machines are effective for efficiently classifying Congressional bills. On some types of bills, the system has difficulty compared to an expert coder. But, in the balance, the algorithm is quite compact and robust. Considering the complexity of coding legislative text into one of 226 subtopics, its effectiveness is about as good as can be expected when using techniques based solely on the “bag of words” principle. Future research should examine using other features which could improve the system as well as other algorithms.

The described algorithm also displays another highly desirable trait for the task – it is easily extensible with additional features. The SVM system is capable of considering out-of-band data to aid in reaching a conclusion in text classification. In concrete terms, the system could be told to consider a count of THOMAS LIV classifications, sponsor committee membership, and other relevant information when predicting the subtopic of a bill. With the correct tools, extending the system to improve its accuracy would then become an exercise for any political science student interested in taking up the task.

The next step for the team is to integrate the algorithm with the human coding team of the Congressional Bills project. Use of the system in their daily work would provide them with the ability to predict the major and subtopic codes for each new Congress' set of bills. Although the system cannot be trusted to generate a 100% accurate answer, it already generates meaningful information useful to understanding when it is making a systemic, likely true prediction versus a wild guess for each bill. This information is critical to the successful adoption of systems like this, and methods to expose this information will be the subject of future research. The team is applying for National Science Foundation funding to pursue these opportunities.

6. ACKNOWLEDGMENTS

Thanks to Dr. John Wilkerson for providing assistance with the Congressional Bills' data. Also, thanks to Dr. Stuart Shulman for encouraging us to submit this document.

7. REFERENCES

- [1] Cristianini, N., Shawe-Taylor, J., and Lodhi, H. Latent semantic kernels. in Brodley, C. and Danyluk, A. *Proceedings of ICML-01, 18th International Conference on Machine Learning*. (San Francisco, US, 2001), Morgan Kaufmann Publishers, pages 66–73.
- [2] Deerwester, S. et al. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407.
- [3] Gwet, K. Kappa Statistic is not Satisfactory for Assessing the Extent of Agreement Between Raters. in *Statistical Methods For Inter-Rater Reliability Assessment*, No. 1, April, 2002.
- [4] Joachims, T. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *Proceedings of the European Conference on Machine Learning (ECML)*. (Springer, 1998)
- [5] Joachims, T. Making Large-Scale SVM Learning Practical. in: *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola (ed.), MIT Press, 1999.
- [6] Kwon, N., Shulman, S.W., and Hovy, E.H.. (Under review). “Collective text analysis for eRulemaking.” *Proceedings of the Sixth National Conference on Digital Government Research*. San Diego, CA.
- [7] Laver, M., Benoit, K., and Garry, J. Extracting policy positions from political texts using words as data. In *American Political Science Review* 97(2).
- [8] Papineni, K. “Why inverse document frequency?” IN *Proceedings of the North American Association for Computational Linguistics, NAACL*, pp. 25–32. (2001)
- [9] Porter, M. F. An algorithm for suffix stripping. Program, 16(3):130–137.
- [10] Sebastiani, F. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1).
- [11] Tokunaga, T. and Iwayama, M. Text categorization based on weighted inverse document frequency. *Technical Report* 94

- TR0001, Department of Computer Science*, (Tokyo Institute of Technology, 1994).
- [12] Yang, H., Callan, J., and Shulman, S. (Under review) “Next steps in near-duplicate detection for eRulemaking.” *Proceedings of the Sixth National Conference on Digital Government Research*. San Diego, CA.
- [13] Yang, Y. and Liu, X. 1999. A re-examination of text categorization methods. In *Proceedings of SIGIR-99*, November.
- [14] Vapnic, V. *The Nature of Statistical Learning Theory*. Springer, New York, NY. 1995.

SESSION 6B

E-RULEMAKING 2

Moderator

José Luis Ambite, University of Southern California, USA

Titles and Authors

Locating Related Regulations Using a Comparative Analysis Approach
Law, Kincho H.; Lau, Gloria T.; Wang, Haoyi

Next Steps in Near-Duplicate Detection for eRulemaking
Yang, Hui; Callan, Jamie; Shulman, Stuart

Progress in Language Processing Technology for Electronic Rulemaking
Shulman, Stuart; Callan, Jamie; Hovy, Eduard; Zavestoski, Stephen

Locating Related Regulations Using a Comparative Analysis Approach

Gloria T. Lau
Research Scientist
Thomson Findlaw
Sunnyvale, CA 94086
glau@stanford.edu

Haoyi Wang
Graduate Student
Stanford University
Stanford, CA 94305-4020
haoyiw@stanford.edu

Kincho H. Law
Professor of Civil and Env. Engr.
Stanford University
Stanford, CA 94305-4020
law@stanford.edu

ABSTRACT

The sheer volume and complexity of government regulations make any attempt to locate, understand and interpret the information a daunting task. Other factors, such as the scattered distribution of the regulations across many sources, different terminologies and cross referencing, further complicate the technical issues in developing a regulation information management system. This paper describes a comparative analysis approach and its potential application to assist locating relevant regulations from different sources. Examples from environmental regulations are employed to illustrate the proposed methodology and framework.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval – retrieval models.

General Terms

Algorithms, Management, Legal Aspects

Keywords

Relatedness Analysis, Regulatory Comparison, Structural Analysis.

1. INTRODUCTION

Regulations are typically specified by Federal as well as State governmental bodies and are often amended by local counties or cities. Regulations emanating from diverse agencies often overlap; because settings and objectives differ they may be difficult to reconcile. As new issues of public safety or fairness arise new regulations are promulgated and must be integrated in the complex existing regulatory framework. The distributed responsibilities for enforcement and compliance assistance increase the complexity of complying with regulations. The scope of concern and the terminology used to express those concerns differs among agencies and industries.

Since environmental regulations have the force of law, it is important that companies be able to locate, understand, and comply with them. It is also advantageous for society to make these regulations as easy to locate and understand as possible so that the environment is protected to the extent provided by the laws in place. However, many have argued that the “complex, evergrowing and oft-adapting ... environmental law is becoming more challenging for practitioners and the judiciary alike” [7]. Furthermore, “there is ample reason to believe that a growing percentage of environmental violations result from a misunderstanding of regulatory requirements or are otherwise unintended” [21].

The burden of complying with environmental regulations can fall disproportionately on small businesses, since these businesses may not have the expertise or resources to keep track of regulations and their requirements [15]. That the requirements of these complex regulations change over time further compounds the problem [21]. As noted in the Washington Post, “Deciphering and complying with federal regulations is a legal and paperwork nightmare for many businesses. To keep pace, some hire consultants to keep track of the applicable health, safety, environmental and equal-opportunity rules” [19]. This burden has been recognized and targeted by legislation designed to address the problem. The Regulatory Flexibility Act (RFA) [16], amended by the 1996 Small Business Regulatory Enforcement Fairness Act (SBREFA) [20], clearly recognizes the information problem facing businesses, particularly small businesses, that must comply with environmental regulations. Although many efforts have been initiated, actual changes in regulation management and dissemination remain a fairly slow process. Advanced ICT technologies and innovative, high quality tools are crucial to further move the regulatory information to the public.

Government regulations are now available on-line but these online portals are primarily designed for displaying information (and often usable only for experienced users). The sheer volume and complexity of this information, coupled with its scattered distribution across many different sources, makes any attempt to locate, understand and interpret the information a daunting task. Some primitive searching capabilities may be provided; however, it remains difficult to locate cross-referenced or related information and to link the information with useful applications, such as compliance assistance. Other factors, such as the high density of inter-referencing within a regulation code and intra-referencing between regulatory documents and the heavy reliance on acronyms, contribute to reducing the readability of the

documents that can be located. Our research objective is to systematically develop formal approaches that will aid locating relevant regulations and assist compliance.

This paper presents a comparative analysis methodology that can be used to search and retrieve regulations as well as to compare regulations from different sources. This paper is organized as follows: Section 2 briefly describes the overall system framework and the development of an XML-based regulatory repository. Section 3 discusses the fundamental methodology employed for comparative analysis of regulations. To illustrate the comparative analysis framework, Section 4 describes example applications in identifying similar provisions from different sources of drinking water standards. Section 5 describes briefly a work-in-progress prototype for locating related provisions across Federal and State regulations. Finally, this paper concludes with a brief discussion on future works in Section 6.

2. SYSTEM FRAMEWORK

The purposes of the regulatory information management (RIM) system are as follows:

- To develop a formal repository to handle diverse regulation files and define a representational structure;
- To develop mechanisms for extracting features and concepts from the regulatory documents and tools for assisting user to identify related regulations;
- To retrieve and compare regulations from different sources on a specific domain topic.

Figure 1 shows the overall framework for the regulatory information management system. There are four basic functions implemented: (1) textual parsing and storage, (2) semi-structured, indexed storage, (3) feature and concept extraction, and (4) comparative analysis and retrieval of related regulatory documents.

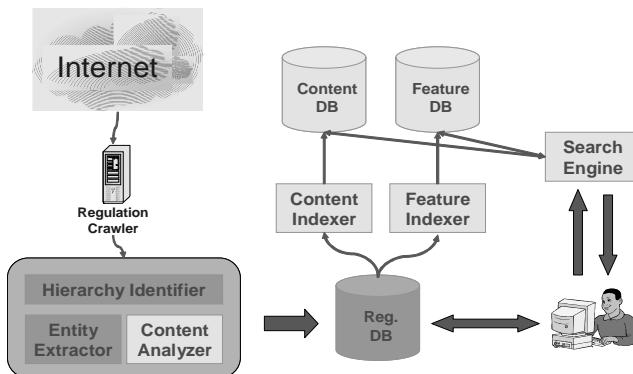


Figure 1. System Framework for the RIM system

To build a repository, the first step is to gather the regulations from diverse sources and transform them into a well-organized XML structure. Since all the State and Federal regulations are now available online, we developed a web crawler to download the raw regulation files from the regulation web sites. Starting with a specific web page containing a regulation, the web crawler is capable of following the links within this web page to further gather other web pages containing the regulations. For each

regulation web site, a configuration file is defined that provides the information about the starting URL, the lowest level the crawlers should traverse, the type of links to follow and the downloadable file types. Figure 2 illustrates the configuration file for retrieving the environmental regulations in the State of Hawaii: the “outputDir” asks the crawler to put downloaded files into directory “HI”; “startTOC” indicates the URL of the first web page the crawler should first traverse; “maxDepth” defines the number of levels of “web pages”, starting from “startTOC”, to be retrieved. For each downloadable web page, the configuration file also defines useful information for traversing the links to find the linked web pages. For example, the variable “linkPattern” defines the pattern to extract useful links from a page; “matchLink” indicates whether the pattern is applied on an embedded link or its anchor text; “filePattern” matches the links to the content; “indexPattern” tells whether a link is an index page. The parameter appended to a variable represents the level number from the start page. Using the web crawler and individually defined configuration file, over a dozen of State regulations have been successfully downloaded and stored in the repository.

```

outputDir = HI
startTOC = http://www.hawaii.gov/dlnr/AdminRulesIdx.htm
maxDepth=2

linkPattern1 = ^Final.*Rules
matchLink1 = false
filePattern1 = .*/dlnr/.*\.\pdf
indexPattern1 = .*

linkPattern2 = *
filePattern2 = .*/dlnr/.*\.\pdf
  
```

Fig. 2: An Example Configuration File for the Web Crawler

```

<regulation id="40.cfr.2" name="PUBLIC INFORMATION" type="federal">
<regElement id="40.cfr.2.A" name="-- Procedures for Disclosure of Records Under the Freedom of Information Act ">
<regText></regText>
<paragraph>
Source: 67 FR 67307, Nov. 5, 2002, unless otherwise noted.
</paragraph>
<regElement id="40.cfr.2.100" name=" General provisions. ">
<regText>
  
```

Figure 3: Example of the XML Structure for Regulation

The downloaded regulations are then transformed into an XML structure, which is well suited for representing semi-structured information. The XML structure is designed to map directly to the hierarchical structure inherent in the regulation documents. For example, we use XML tag “regElement” to label a section in the regulation hierarchy. Each “regElement” in XML file may have a parent and/or multiple children elements representing the corresponding sections and subsections. Another advantage of using XML is that metadata can be added easily to the content. To transform the downloaded regulation content (which are typically in HTML, PDF or WORD format), we first convert the

file into simple texts, if necessary, using utility tool such as XPDF. A shallow parser, which is written in Perl language, is then employed to transform the text into an XML structure as shown in Figure 3. The hierarchical structure of regulations is preserved by properly structuring provisions as XML elements. For instance, Section 40.cfr.2.A is a provision in Section 40.cfr.2, and is thus structured to be a child node of the XML element of Section 40.cfr.2. With the hierarchical organization captured in the XML structure, rendering tools can easily be developed to display and view the regulations in its natural organization.

3. COMPARATIVE ANALYSIS

The proliferation of the Internet has led to an extensive amount of research on retrieving relevant documents based on keyword search [2]. Well-established techniques such as query expansions [8, 17] have been deployed to increase retrieval accuracy, with a significant amount of subsequent developments [1, 6, 14, 22] to improve performance. Thus, most repositories are equipped with a search and browse capability for viewing and retrieval of documents. It is reasonable to assume the following in a regulatory repository: at least one relevant document will be located by the user either with keyword search or by browsing through an ontology. In this section, we will discuss the techniques we use to suggest to the users similar provisions from different sources of regulations, starting from a correctly identified section. In essence, we focus on refining the back end comparison technique for documents based on a deep understanding of regulations, taking advantage of domain knowledge and structural organizations, rather than matching queries at the front end [11].

Since a typical regulation can easily exceed thousands of pages, a comparison between a full set of regulation and another is meaningless [3]. Instead, a section from one set of regulation is compared with another section from another set, such as a comparison between Section 141.32.e.16 in Code of Federal Regulations Title 40 (40CFR) [5] and Section 64468.1(c) in California Code of Regulations Title 22 (22CCR) [4]. The analysis computes a similarity score, which measures the *degree of similarity* between two documents. The score is defined on a relatedness measurement interval that ranges from 0 to 1, with 0 representing unrelated materials and 1 being the most related or identical materials. The similarity score is denoted by $f(F, C) \in [0, 1]$ per pairs of provisions, for example, pair (F, C) with Section F from the Federal standard 40CFR and Section C from the California code 22CCR. Naturally, the comparison is commutative. In other words, we have $f(F, C) = f(C, F)$.

The similarity score represents the direct content comparison of provisions based on different feature matching. Feature is the evidence of relatedness between two provisions, which could contain domain-specific information. There are generic features that are common across all domains of regulations, such as exceptions, definitions and concepts. For instance, examples of concepts found in drinking water regulations are “ground water” and “levels of exposure.” The second type of features are domain-specific ones, such as glossary terms defined in engineering handbooks, author-prescribed indices at the back of reference books, measurements found in prescriptive regulations, and chemicals and effective dates specific to environmental regulations.

The similarity score between two sections is computed as a linear combination of the scores obtained using different feature matching, which allows for a combination of generic features, such as concepts, as well as domain knowledge, such as drinking water contaminants in environmental regulations. This design provides the flexibility to add on features and different weighting schemes if domain experts desire to do so. The scoring scheme for each of the features essentially reflects how much resemblance can be inferred between the two sections based on that particular feature. For instance, concept matching is done similar to the index term matching in the Vector model [18], where the degree of similarity of documents is evaluated as the correlation between their index term vectors. Using this Vector model, we take the cosine similarity between the two concept vectors as the similarity score based on a concept match. We use only the frequency count as the concept weight and deliberately exclude any inverse document frequency (*idf*) component here. This is because concepts represent an already selected set of important noun phrases which do not include common terms such as stopwords, and therefore no *idf* factor is included in our model.

Here, our usage of the Vector model differs from generic applications in two ways. Our comparison is on extracted features, such as measurements, but not index terms; in addition, we have a much more selective collection of documents, namely regulations in certain domains rather than a general-purpose corpus. If one desires to incorporate domain knowledge, axis independence no longer holds. For instance, some features are characterized by ontologies to define synonyms. Some features simply cannot be modeled as Boolean term matches due to their inherent non-Boolean property, such as measurements, (As an example, a domain expert can potentially define a measurement of “12 inches maximum” as 75% similar to a measurement of “12 inches.”) Some domain-specific features are supplemented with feature dependency information defined by knowledge experts, who do not necessarily agree with a Boolean definition. It is unrealistic to assume that the world can be modeled as a Boolean match, and as a result, domain knowledge is potentially non-Boolean. In essence, the degree of match between two features is no longer limited to only 0% or 100%.

To accommodate a non-Boolean degree-of-match algorithm, we propose a vector space transformation based on the Vector model. For features with defined synonyms or a non-Boolean matching scheme, the feature vectors are mapped onto a different vector space before a cosine comparison. A linear transformation in the form of $\vec{m}' = D\vec{m}$, where D denotes the transformation matrix, is employed to account for axis dependencies introduced by user-defined partial match algorithms. In other words, D captures available domain knowledge, and projects the feature vector \vec{m} onto an alternate space where the resultant vector $\vec{m}' = D\vec{m}$ represents the consolidated feature frequencies. Details and proofs of the formulation are given in [10]. The transformation is shown to produce consistent results when synonymous information are modeled using two different spaces, namely the original n -dimensional space and a reduced vector space with the synonymous feature axes collapsed into one.

4. SIMILAR PROVISIONS FROM DIFFERENT SOURCES – EXAMPLE FROM DRINKING WATER STANDARDS

To illustrate the use of domain knowledge such as ontological information and the associated vector space transformation, we will discuss one particular example of feature extraction and matching here. We focus on drinking water standards in environmental regulations, where certain chemicals play an important role in this domain. In particular, the US Environmental Protection Agency (EPA) publishes an index of national primary drinking water contaminants [13]. This list contains about a hundred potential drinking water contaminants; examples include “trans-1,2-dichloroethylene,” “vinyl chloride” and so on.

An ontology is developed based on the index of drinking water contaminants published by the EPA as well as supplementary materials, and an excerpt is shown in 4. A category name is preceded by an exclamation mark, while elements belonging to the category are signaled with a plus sign. For instance, a domain expert can easily codify synonymous / acronymic information such as “total trihalomethane” and “tthm” as shown in the ontology. This further illustrates the need to incorporate domain knowledge, where most intelligent mining tools are likely to fail to identify such type of information even with the help of a dictionary¹.

```
!Disinfectants and Disinfection-byproducts
!Disinfectants
...
!Chlorine
  +chlorine
  +cl2
  +hypochlorite
  +hypochlorous acid
!Disinfection Byproducts
  +d/dbp
  +d/dbps
  +dbp
  +dbps
...
!Total Trihalomethanes
  +trihalomethane
  +tthm
  +tthms
...
```

Figure 4: Ontology Developed on Drinking Water Contaminants

To incorporate this piece of domain knowledge, our XML parser takes the ontology as a flat list and tags the drinking water contaminants as `<dwc>` subelements in provisions where they appear. Other features, such as `<measurement>`, are tagged using handcrafted rules that match general patterns in regulations. The rules are implemented specifically for and after a careful study of regulations in our working domain. Possibly due to the

¹ In this particular example, the term “tthm” cannot be found in either Webster or Oxford dictionary. Merriam-Webster Collegiate Dictionary is a product of Merriam-Webster, Inc.; Oxford English Dictionary is a product of Oxford University Press.

rigorous nature of regulation drafting, the pattern matching rules for `<measurement>` seem to perform well. We have yet to uncover any mismatches in the sections that we have reviewed, but a formal evaluation is in due course.

As shown in Figure 5, stemming and frequency counting for `<dwc>` elements are performed as in the `<concept>` feature. The difference between a `<concept>` element and a `<dwc>` element is that `<concept>` is a general key phrase without any associated domain knowledge, whereas other elements, for instance, `<dwc>` as shown below, are defined with domain knowledge such as an ontology.

```
<dwc name="arsen" times="1" />
```

The terms or phrases, such as “arsenic”, might be extracted as a concept already, however its sheer presence in the dwc list adds to its importance in this particular domain. Using the ontological information as shown in Figure 4, feature matching can now identify important vocabularies in the domain of drinking water regulations. Similarity computation is enhanced with synonymous information such as phrases like “disinfection byproducts” and “d/dbp”. The transformation matrix D would represent synonymous information in the ontology in this example, and the consolidated frequency vector \vec{m} would contain the consolidated frequency counts of synonyms. The similarity computation would count the frequency of “disinfection byproducts” combined with “d/dbp” on the same feature axis.

Original section 141.11.b from the 40 CFR § 141.11 Maximum contaminant levels for inorganic chemicals.

(a) The maximum contaminant level for arsenic applies only to community water systems ...

(b) The maximum contaminant level for arsenic is 0.05 milligrams per liter for community water systems until January 23, 2006.

Refined section 141.11.b in XML format

```
<regElement id="40.cfr.141.11.b" name="">
  <dwc name="arsen" times="1" />
  <concept name="commun water system" times="1" />
  <measurement unit="mg/l" size="0.05" quantifier="max" />
  <date to="January 23, 2006" num="1" />
  ...
  <regText>
    The maximum contaminant level for arsenic is 0.05 milligrams per liter for community water systems until January 23, 2006.
  </regText>
</regElement>
```

Figure 5: Drinking Water Contaminant and Effective Date Tags

As a result of the similarity analysis, related provisions can be retrieved and recommended to users based on the resulting scores. Different combinations of features, different feature weights, and different feature scoring schemes can be experimented.

Preliminary results are shown using an equal weight of concepts, measurements, drinking water contaminants, and effective dates in the domain of drinking water regulations.

Our comparison system is tested on different groups of regulations, such as comparisons among accessibility regulations, comparisons among drinking water standards, and cross domain comparisons. The average similarity scores among drinking water regulations are relatively small compared to that of accessibility regulations, possibly because of the volume and diversity of coverage of drinking water regulations. Comparing different features among drinking water standards, similarity appears to be captured mostly by concepts. This is understandable since terms form the basis of body text in regulations, and thus appear much more often than non term-based features such as measurements. Other term-based features, such as drinking water contaminants, also result in average similarity scores bigger than those obtained using other features such as measurements. Overall, the use of an ontology to help identify synonyms seems to boost the retrieval of similar sections. Effective dates and measurements are comparatively less significant, possibly reflecting on the fact that they are non term-based features and the scoring schemes are more unsparing than that of drinking water contaminants or concepts.

Two examples are given to illustrate the similarity and

dissimilarity between Federal and State drinking water regulations. The first example, shown in Figure 6, is a top ranked pair of related provisions on drinking water control of the chemical Barium required by the 40CFR and 22CCR. This pair of provisions is actually identical in text except the subject of governing agency changes between Environmental Protection Agency (EPA) and California Department of Health Services (DHS). It is not uncommon that one agency directly adopts provisions issued by another agency.

In this example of Barium requirements, the text in the provision is actually somewhat unusual and does not seem to be written in standard regulatory language. The text appears to be a *notice* required by both the EPA and the California DHS, where the notice could potentially come from an outside source. The careful reader might also note that the EPA and the California DHS *do* have different Barium requirements – the EPA requires 2 parts per million while the California DHS sets the requirement at 1 part per million. It appears that the two agencies might have modified the notice according to their separate standards. This example also illustrates the importance of domain knowledge, where a measurement comparison would reveal that these two provisions are not identical, even though the wordings are almost the same.

These near-identical provisions are difficult to locate for the human eyes; without extracting the measurement feature, an

Code of Federal Regulations Title 40

141.32.e.16 Barium

The United States Environmental Protection Agency (EPA) sets drinking water standards and has determined that barium is a health concern at certain levels of exposure. This inorganic chemical occurs naturally in some aquifers that serve as sources of ground water. It is also used in oil and gas drilling muds, automotive paints, bricks, tiles and jet fuels. It generally gets into drinking water after dissolving from naturally occurring minerals in the ground. This chemical may damage the heart and cardiovascular system, and is associated with high blood pressure in laboratory animals such as rats exposed to high levels during their lifetimes. In humans, EPA believes that effects from barium on blood pressure should not occur below 2 parts per million (ppm) in drinking water. EPA has set the drinking water standard for barium at **2 parts per million (ppm)** to protect against the risk of these adverse health effects. Drinking water that meets the EPA standard is associated with little to none of this risk and is considered safe with respect to barium.

California Code of Regulations Title 22

64468.1(c) Barium

The California Department of Health Services (DHS) sets drinking water standards and has determined that barium is a health concern at certain levels of exposure. This inorganic chemical occurs naturally in some aquifers that serve as sources of ground water. It is also used in oil and gas drilling muds, automotive paints, bricks, tiles and jet fuels. It generally gets into drinking water after dissolving from naturally occurring minerals in the ground. This chemical may damage the heart and cardiovascular system, and is associated with high blood pressure in laboratory animals such as rats exposed to high levels during their lifetimes. In humans, DHS believes that effects from barium on blood pressure should not occur below 2 parts per million (ppm) in drinking water. DHS has set the drinking water standard for barium at **1 part per million (ppm)** to protect against the risk of these adverse health effects. Drinking water that meets the DHS standard is associated with little to none of this risk and is considered safe with respect to barium.

Figure 6: Direct Adoption of Provisions Across Federal and California State on the Topic of Drinking Water Standards

automated comparison using bag-of-word type of analysis will return a similarity score close to 1. The identification of related provisions brings us a step closer to the identification of different provisions on the same topic. Based on our framework, a potential improvement can be envisioned to capture minor differences between provisions. Assuming that the interested provisions are related, we first apply the relatedness analysis system to identify the most related pairs of provisions, such as the requirements on Barium by the California and Federal agencies. Different features, such as measurements, can be compared individually to capture differences between provisions. This will require a formal definition and formulation of a difference operator between provisions.

Aside from adopting identical provisions between Federal and State agencies, differences are also observed between the two documents. For instance, the 40CFR makes use of many chemical acronyms, such as TTHM, whereas the full term “total trihalomethanes” is always spelled out in the 22CCR. Figure 7 shows a pair of provisions illustrating the case. Based on a pure concept match, the two provisions result in zero similarity. The similarity score based on a drinking water contaminant match is 0.49, due to the use of ontological information as shown in Figure 7 that identifies the acronym TTHM as a match to “total trihalomethanes,” as well as HAA with “haloacetic acids.” This example justifies for the incorporation of domain knowledge; without which, a user searching for TTHM or HAA will fail to find anything in 22CCR but only in 40CFR.

To show the dissimilarity between different domains of regulations, we compared drinking water standards 40CFR with fire protection standards in Chapter 9 of the International Building Code (IBC) [9]. All of the features but concepts show a zero similarity score. Features such as drinking water contaminants and effective dates only exist in environmental regulations, which explains why the fire code does not share any of them. Both domains contain measurements; however, they are very different kinds of measurements that are not shared between the two domains, such as “75 feet clearance” in the fire code and “2 parts per million” in drinking water standards. Concepts generate a close-to-zero similarity score, as there are still some common phrases that are shared, such as the phrase “common area” found in both domains.

One example is shown in Figure 8, where provisions from the two separate domains share some remote similarity. Section 141.85.a.1.iv.B.6 from the 40CFR is a small subsection under Section 141.85 on “public education and supplemental monitoring requirements.” This section happens to touch on the safety of *electrical systems* in public education. Section 907.2.8.1 from the IBC deals with fire detection systems that involves discussion of *electrical systems* as well. These two tangentially related provisions that are top ranked among this group of cross-domain comparisons are one of the few related provisions found by our system with negligible similarity scores.

Code of Federal Regulations Title 40

141.132.a.2 [No Title; under Monitoring Requirements]

Systems may consider multiple wells drawing water from a single aquifer as one treatment plant for determining the minimum number of **TTHM** and **HAA5** samples required, with State approval in accordance with criteria developed under §142.16(h) (5) of this chapter.

California Code of Regulations Title 22

64823(e) [No Title; under Field of Testing]

Field of Testing 5 consists of those methods whose purpose is to detect the presence of trace organics in the determination of drinking water quality and do not require the use of a gas chromatographic/mass spectrophotometric device and encompasses the following Subgroups: EPA method 501.1 for trihalomethanes; EPA method 501.2 for trihalomethanes; EPA method 510 for **total trihalomethanes**; EPA method 508 for chlorinated pesticides; EPA method 515.1 for chlorophenoxy herbicides; EPA method 502.1 for halogenated volatiles; EPA method 503.1 for aromatic volatiles; EPA method 502.2 for both halogenated and aromatic volatiles; EPA method 504 for EDB and DBCP; EPA method 505 for chlorinated pesticides and polychlorinated biphenyls; EPA method 507 for the haloacids; EPA method 531.1 for carbamates; EPA method 547 for glyphosate; EPA method 506 for adipates and phthalates; EPA method 508A for total polychlorinated biphenyls; EPA method 548 for endothall; EPA method 549 for diquat and paraquat; EPA method 550 for polycyclic aromatic hydrocarbons; EPA method 550.1 for polycyclic aromatic hydrocarbons; EPA method 551 for chlorination disinfection byproducts; EPA method 552 for **haloacetic acids**.

Figure 7: Terminological Differences Between Federal and State Regulations on the Topic of Drinking Water Standards

Code of Federal Regulations Title 40

141.85.a.1.iv.B.6 [No title; under Public Education and Supplemental Monitoring Requirements]

Have an electrician check your wiring. If grounding wires from the **electrical system** are attached to your pipes, corrosion may be greater. Check with a licensed electrician or your local electrical code to determine if your wiring can be grounded elsewhere. DO NOT attempt to change the wiring yourself because improper grounding can cause electrical shock and **fire** hazards.

International Building Code, Chapter 9

907.2.8.1 Fire Detection System

System smoke detectors are not required in guestrooms provided that the single-station smoke alarms required by Section 907.2.10 are connected to the emergency **electrical system** and are annunciated by guestroom at a constantly attended location from which the **fire** alarm system is capable of being manually activated.

Figure 8: Remotely Related Provisions Identified from a Drinking Water Regulation and a Fire Code

5. LOCATING SIMILAR REGULATIONS

As demonstrated in the previous sections, the knowledge-driven comparative analysis approach is potentially capable of discovering similar regulatory provisions. In an ongoing work, we are extending the methodology and framework to develop a “regulatory locator” for domain specific applications. While regulations on a specific domain are mostly grouped under a specific title or part(s), related regulations also exist in other titles or parts. For example, “mercury”, a specific chemical, which appears in 40.CFR.141 (Part 141 of Title 40 of CFR), also appears in Title 21 of CFR on Drug and Food Administration. Similarly, the term “mercury”, which appears mostly in Title 22 of CCR, also appears in Titles 17 (Public Health), 8 (Industrial Relations), 3 (Food and Agriculture) and other parts of CCR. To fully locate “all” regulations for a specific hazardous substance is a very difficult task. Current attempts to classify industry related regulations are mainly done manually. Our objective is to apply the comparative analysis framework and study the feasibility of extending the tool to facilitate the development of “regulatory locators” (RegLocator) for different domain specific application.

To enable users to quickly search and find regulations of interests, a search system that utilize terms, concepts and structural relationships, has been implemented. Figure 9 shows the GUI for the RegLocator prototype system. The user can define the primary source of interest for finding the regulations for a particular subject. At the same time, the user can also specify a secondary source to locate possibly related regulations. For instance, if the user is interested in “waste water” in the Federal code, the user can also find related regulations from the secondary source, such as California. This feature is implemented through both a search mechanism and the comparative analysis system. Currently, the environmental codes (which were obtained using the web crawler and text parser discussed earlier) in the RegLocator repository include the Federal (CFR) and three States (Alabama, Arizona and California) regulations.

Figure 10 shows the search results from the query “waste water” from the Federal regulations. Additionally, a set of related (domain) concepts are also shown that also indicate their “relevance” with the query’s key words. The related concepts could potentially be useful for appending the terms to the previous query to form a new query or for providing hints when issuing a new query. From the search results, the user can browse and retrieve a specific provision of interest (see Figure 11). Besides the text of the provision, other provisions located in the “vicinity” of the retrieved provision are also shown in a hierarchical structure (which reflects the typical structure of a regulatory document). Furthermore, related concepts and terms for the provision are also shown to enable searching for further results. Last but not least, related regulations from the secondary source are also shown. User can then search and browse the related regulations, possibly for comparison purposes.

6. DISCUSSION

In developing a regulation information management (RIM) system that would allow searching, retrieving and comparing regulations, domain knowledge plays a very important role in understanding regulations and the relationships between them. We believe a knowledge driven approach, combining with similar analysis, is a powerful way to develop the RIM system. In particular, distinct knowledge sources or regulations do not have to be made completely consistent, only the terms and the concepts that *articulate* their application connections are involved. In this study, we demonstrated the use of an ontology to match features in drinking water standards. With our current XML repository of environmental regulations from the Federal and States, the RegLocator can be enhanced to incorporate domain knowledge to help retrieval of related provisions from different states as well as matching keywords. For instance, users typing in “disinfection byproduct” will now be able to locate provisions written using the acronym “dbp.” To this end, we plan to collect and to develop, by

Regulation Search Page - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://171.64.55.140/regsearch/

Customize Links Free Hotmail Windows Media Windows

Regulations' Boolean Query Interface

Regulation Sources: Federal AL AZ CA

Related Regulations: Federal AL AZ CA

Search Tutorial:

- All three input fields are connected by boolean operator "AND".
- In each field, you may input phrases, terms, and prefixes.
 - A phrase is multiple words delimited by space, like "waste water";
 - A term is a single word, like "mercury";
 - A prefix is a term with a "*" at end, like "env*";
 - The search items are separated by ",", like "waste water, mercury, env*".
- The search items can be connected by "AND" (Include all words), "OR" (Include any word), and "NOT" (Exclude all words).
- The default connection is "AND" for input "Content" and "Title", and is "OR" for input "ID".

Regulation Content: include all words

(e.g. waste, water, EPA)

Regulation Title: include all words

(e.g. air, waste)

Regulation ID: (e.g. 40.cfr.273.53.a)

Figure 9: GUI for RegLocator – A Prototype System for Regulation Locator

Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://171.64.55.140/regsearch/search?resource=FED&reference=C&content=waste+water&content_blop=AND&title=thane_blop=AND&id=1&start=0

Customize Links Free Hotmail Windows Media Windows

Page 1

Result pages: 1 2 3 4 5 6 7 8 9 10 next

FED 40 CFR 428.80 Applicability; description of the wet digestion reclaimed rubber subcategory.
The provisions of this subpart are applicable to process waste water discharges resulting from the production of reclaimed rubber by use of the wet digestion process.

FED 40 CFR 428.90 Applicability; description of the pan, dry digestion, and mechanical reclaimed rubber subcategory.
The provisions of this subpart are applicable to process waste water discharges resulting from the production of reclaimed rubber except when produced by the wet digestion process.

FED 40 CFR 240 204 Recommended procedures; Design.
40 CFR 240 204 effluent waters should not be discharged indiscriminately. Consideration should be given to onsite treatment of process and waste waters before discharge. 40 CFR 240 204 bRecirculation.

FED 40 CFR 412 15 Standards of performance for new sources.
40 CFR 412.15 aSubject to the provisions of paragraph (b) of this section, the following standards of performance establish the quantity or quality of pollutants or pollutant properties which may be...

FED 40 CFR 412 15 Effluent limitations guidelines representing the degree of effluent reduction attainable by the application of the best available technology economically achievable.
40 CFR 412.13 aSubject to the provisions of paragraph (b) of this section, the following limitations establish the quantity or quality of pollutants or pollutant properties which may be discharged by...

FED 40 CFR 428 110 Applicability; description of the latex foam subcategory.
The provisions of this subpart are applicable to process waste water discharges resulting from the manufacture of latex foam except for those discharges from textile plants subject to the provisions...

FED 40 CFR 409 72 Effluent limitations guidelines representing the degree of effluent reduction attainable by the application of the best practicable control technology currently available.
Except as provided in Section 125.30 through 125.32, and subject to the provisions of paragraph (a) of this section, any existing point source subject to this subpart shall achieve the...

FED 40 CFR 427 61 Specialized definitions.
For the purpose of this subpart 40 CFR 427.61 aExcept as provided below, the general definitions, abbreviations and methods of analysis set forth in 40 CFR part 401 shall apply to this...

FED 40 CFR 409 13 Effluent limitations guidelines representing the degree of effluent reduction attainable by the application of the best available technology economically achievable.
40 CFR 409.13 aThe following limitations establish the quantity or quality of pollutants or pollutant properties which may be discharged by a point source where the sugar beet processing capacity of...

FED 40 CFR 436 31 Specialized definitions.
For the purpose of this subpart 40 CFR 436.31 aExcept as provided below, the general definitions, abbreviations and methods of analysis set forth in part 401 of this chapter shall apply to this...

Done

Figure 10: Search Results and Related Concepts

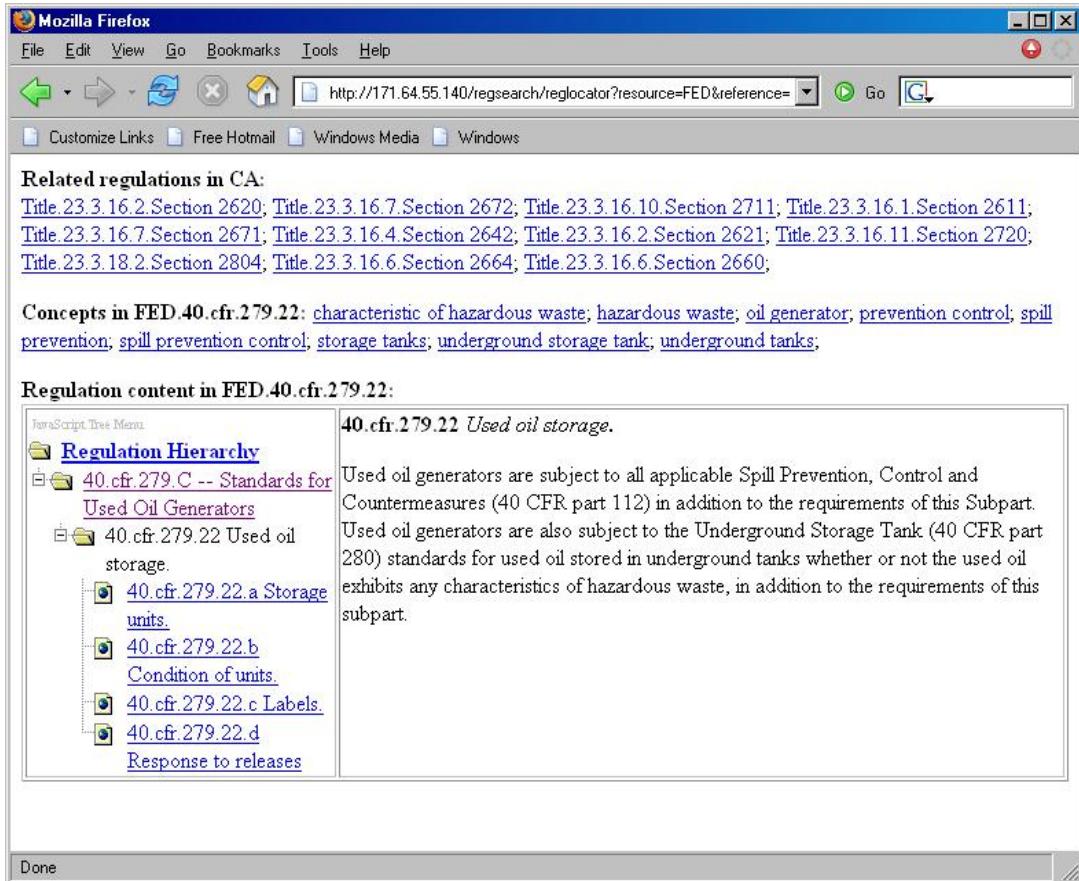


Figure 11: Regulation Displayed and Related Provisions in Secondary Source

way of collaboration with industry experts, ontological information relating to other sub-domains within environmental regulations. We also plan to study and to implement ontological composition [12] once a satisfactory set of ontologies is developed.

We have partially evaluated the performance of the similarity analysis system by comparing results from our system to that of a traditional retrieval system. Preliminary study shows that our system outperformed a traditional index term analysis, especially with the use of domain knowledge [10,11]. A formal evaluation is planned to estimate the precision and recall of the RegLocator system. However, due to the size of the regulatory repository and the complexity of the law, we plan to scope the evaluation to drinking water standards in a few selected states. A traditional bag-of-word Vector model will serve as the baseline, and we will explore different combinations of related concepts and related provisions to improve the accuracy of a keyword search.

7. ACKNOWLEDGMENTS

This research project is sponsored by the National Science Foundation, Grant Numbers EIA-9983368 and EIA-0085998. The authors would like to thank Mr. Bill Labiosa for developing the ontology for the drinking water contaminants. The authors

would also like to acknowledge an equipment grant from Intel Corporation.

8. REFERENCES

- [1] R. Attar and A.S. Fraenkel. "Local Feedback in Full-Text Retrieval Systems," *Journal of the ACM*, 24 (3), pp. 397-417, 1977.
- [2] M.W. Berry and M. Browne. *Understanding Search Engines: Mathematical Modeling and Text Retrieval*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1999.
- [3] L.K. Branting. "Reasoning with Portions of Precedents," In *Proceedings of the 3rd International Conference on Artificial Intelligence and Law (ICAIL 1991)*, Oxford, England, pp. 145-154, June 25-28, 1991.
- [4] *California Code of Regulations (CCR)*, Title 22, California Office of Administrative Law, Sacramento, CA, 2003.
- [5] *Code of Federal Regulations (CFR)*, Title 40, Parts 141 - 143, US Environmental Protection Agency, Washington, DC, 2002.
- [6] C.J. Crouch and B. Yang. "Experiments in Automatic Statistical Thesaurus Construction," In *Proceedings of the 15th Annual International ACM SIGIR Conference on*

- Research and Development in Information Retrieval*, Copenhagen, Denmark, pp. 77-88, 1992.
- [7] E. L. Dawson and L.L. Davies, "Book Review: Environmental Law And Policy: Nature, Law, And Society. By Zygmunt J.B. Plater, Robert H. Abrams, William Goldfarb, And Robert L. Graham," *Stanford Environmental Law Journal*, Volume 19, Number 2, pp. 469-478, May 2000.
- [8] E. Ide. "New Experiments in Relevance Feedback," In G. Salton (Eds.), *The SMART Retrieval System - Experiments in Automatic Document Processing*, Prentice Hall, Inc., Englewood Cliffs, NJ, 1971.
- [9] *International Building Code 2000*, International Conference of Building Officials (ICBO), Whittier, CA, 2000.
- [10] G. Lau. *A Comparative Analysis Framework for Semi-Structured Documents, with Applications to Government Regulations*, Ph.D. Thesis, Civil and Environmental Engineering, Stanford University, Stanford, CA, 2004.
- [11] G. T. Lau, K. H. Law, and G. Wiederhold. "A Relatedness Analysis of Government Regulations using Domain Knowledge and Structural Organization," (accepted for publication) *Information Retrieval*.
- [12] P. Mitra and G. Wiederhold, "Resolving Terminological Heterogeneity in Ontologies," *Proceedings of Workshop on Ontologies and Semantic Interoperability at the 15th European Conference on Artificial Intelligence (ECAI)*, Lyon France, 2002.
- [13] *Potential Drinking Water Contaminant Index*, US Environmental Protection Agency, Washington, DC, 2003.
- [14] Y. Qiu and H.-P. Frei. "Concept Based Query Expansion," In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Pittsburgh, PA, pp. 160-169, 1993.
- [15] C. Rechtschaffen, "Competing Visions: EPA And The States Battle For The Future Of Environmental Enforcement," *Environmental Law Reporter*, 30 Envtl. L. Rep. 10803, September 2000.
- [16] *Regulatory Flexibility Act (RFA)*, 5 U.S.C. §§ 601 et seq, 1980.
- [17] J.J. Rocchio. "Relevance Feedback in Information Retrieval," In G. Salton (Eds.), *The SMART Retrieval System - Experiments in Automatic Document Processing*, Prentice Hall, Inc., Englewood Cliffs, NJ, 1971.
- [18] G. Salton. *The Smart Retrieval System - Experiments in Automatic Document Processing*, Prentice Hall, Englewood Cliffs, NJ, 1971.
- [19] C. Skrzyncki, "The Regulators; Compliance Education Goes Self-Service", The Washington Post, May 23rd, 2000.
- [20] *Small Business Regulatory Enforcement Fairness Act (SBREFA)*, Pub Law No. 104-121, March 29 1996 (available at <http://www.epa.gov/sbrefa/statute.htm>).
- [21] D. B. Spence, "Paradox Lost: Logic, Morality, and the Foundations of Environmental Law in the 21st Century," *Columbia Journal of Environmental Law*, Volume 20, Issue 1, pp. 145-182, 1995.
- [22] J. Xu and W.B. Croft. "Query Expansion Using Local and Global Document Analysis," In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Zurich, Switzerland, pp. 4-11, 1996.

Next Steps in Near-Duplicate Detection for eRulemaking

Hui Yang

Language Technology Institute
School of Computer Science
Carnegie Mellon University

+1-412-268-4083

huiyang@cs.cmu.edu

Jamie Callan

Language Technology Institute
School of Computer Science
Carnegie Mellon University

+1-412-268-4525

callan@cs.cmu.edu

Stuart Shulman

Library and Information Science
School of Information Sciences
University of Pittsburgh

+1-412-624-3776

shulman@pitt.edu

ABSTRACT

Large volume public comment campaigns and web portals that encourage the public to customize form letters produce many near-duplicate documents, which increases processing and storage costs, but is rarely a serious problem. A more serious concern is that form letter customizations can include substantive issues that agencies are likely to overlook. The identification of exact- and near-duplicate texts, and recognition of unique text within near-duplicate documents, is an important component of data cleaning and integration processes for eRulemaking.

This paper presents DURIAN (**D**UPLICATE **R**EMOVAL **I**n **lA**rge collectioN), a refinement of a prior near-duplicate detection algorithm DURIAN uses a traditional bag-of-words document representation, document attributes ("metadata"), and document content structure to identify form letters and their edited copies in public comment collections. Experimental results demonstrate that DURIAN is about as effective as human assessors. The paper concludes by discussing challenges to moving near-duplicate detection into operational rulemaking environments.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval] Clustering, Query formulation, Retrieval models, Search process

General Terms

Algorithms, Performance, Experimentation.

Keywords

Duplicate detection, clustering, eRulemaking, public comments, information retrieval, text analysis

1. INTRODUCTION

U.S. law and standard regulatory practice requires U.S. regulatory agencies to give notice of a proposed rule and then respond to substantive comments from lobbyists, companies, trade organizations, special interest groups, and the general public before issuing a final regulation or rule [5][12][18]. When the comment volume is low, as is usually the case, this task is a minor burden. However, a small number of high profile regulations can attract hundreds of thousands of comments, most of which are exact or near duplicate form letters [7][21]. Near duplicates are generated in large volume when an organization posts a form letter on a Web page that allows or encourages customization before submission. For example, letters and email generated by

MoveOn.org for a recent rule were characterized by two form letter sentences followed by text ranging in length from one sentence to, on a few rare occasions, many paragraphs.

Currently, simple heuristics are used to manually identify the many copies of form letters. Often this work is conducted by consulting firms, rather than agency personnel. Most form letter customizations express largely the same opinion in slightly different language [20], but occasionally customizations include substantive issues, which the agencies or their contractors are likely to overlook [19]. The automatic sorting of documents into exact and near duplicate categories, as well as the automatic identification of modified passages, could significantly lower the costs and risks involved in processing large volumes of public comments.

One novel characteristic of near-duplicate detection for notice and comment rulemaking is that two comments may be considered near-duplicates even if they share a relatively small amount of text. A public interest group may encourage people to personalize a form letter by appending their own text to text written by the interest group. A regulatory agency would likely want these comments grouped together even though the amount of modified text varied greatly.

Early research on duplicate detection was done mostly in the areas of databases, digital libraries, and electronic publishing. Recently, duplicate detection has been studied for web search tasks, for example, to give more effective and efficient web-crawling, document ranking, and document archiving. Duplicate detection techniques have been proposed that range from manually coded rules to applications of the latest machine learning techniques [1][2][6][10][14][15][17][22]. Their focus varies from providing high detection rates to minimizing the computational and storage resources. Accuracy varies as well. For large collections, some techniques are too expensive computationally to be deployed in their full capacity. Some algorithms are very efficient yet very brittle and sensitive to even small changes of the text.

This paper proposes a similarity measure for duplicate detection that is based on Kullback-Leibler (KL) distance. It also investigates the use of clustering techniques to find near-duplicate documents. Data clustering is a popular approach for automatically grouping similar objects. In practice this discovery process should avoid redundancies with existing knowledge about groupings, and reveal novel, previously unknown aspects of the data. The technique proposed below uses instance-level clustering

constraints based on document attributes and the editing styles of near-duplicates: *Block Edit* (add or delete several paragraphs), *Key Block* (contains one or more well-known paragraphs), *Minor Change* (small editing changes), *Minor Change & Block Edit Combination*, and *Block Reordering* (reorder known paragraphs). We show that an existing clustering algorithm can be modified to enforce these constraints.

Our goal in this work is to greatly improve near-duplicate detection accuracy for notice and comment rulemaking as well as to maintain efficiency. Our methods are evaluated in experiments with subsets of a public comment database collected for a recent U.S. Environmental Protection Agency (EPA) rulemaking. The experimental results show that system-human intercoder agreement is comparable to human-human intercoder agreement.

The rest of the paper is organized as follows. Section 2 details our algorithm. Section 3 describes our evaluation methodology. Section 4 presents experimental results. Section 5 discusses the next steps along this research path, and concludes.

2. ALGORITHM OVERVIEW

Our goal is a software system that assists rule writers and other interested parties in understanding comments that U.S. regulatory agencies receive as part of notice and comment rulemaking. The system should identify reference copies of form letters, modified copies of form letters (near-duplicates) and how they were modified, and unique comments. The system should make decisions that are consistent with human assessments in this domain, which means that documents might be considered near-duplicates even if they have only relatively small passages in common. Duplicate and near-duplicate detection for notice and comment rulemaking in the era of pervasive email and the Web has several characteristics that help define or constrain the problem. The sections below describe an algorithm that meets these requirements.

2.1 Previous Algorithm

An earlier version of our research [22] uses a two stage duplicate detection algorithm. The first stage involves feature-based retrieval. The second stage uses a single-pass clustering algorithm to group the near duplicates.

The system starts with lexical preprocessing intended to normalize the representation of documents. Document mark-up such as HTML tags and email headers are automatically removed. Metadata (email senders, relayers, and recipients; timestamps), address blocks, and signature blocks are recognized and tagged automatically, using simple, rule-based heuristics.

Any comment that has more than 5 exact duplicates (after lexical preprocessing) is considered an instance of a form letter. The copy with the earliest timestamp is the *reference copy* of the form letter. Each reference copy becomes a *seed document* for the clustering algorithm. Some documents that are not reference copies may be also become seeds later in the clustering process.

The body text of each seed document is broken into chunks. Chunks are constrained to not cross paragraph boundaries. The number of chunks created from a document is determined by a heuristic step function. The number of words in each chunk is:

$$\begin{aligned} \text{if document length} > N : & m \\ \text{otherwise :} & \text{document length / } n \end{aligned} \quad (1)$$

If a document contains more than N words, the size of each chunk is set to m words; if it contains fewer words, there will be at most n chunks within it. Thus the compression ratio is higher for longer documents and lower for short ones. In our system, we set $N=200$, $m=40$, $n=8$ empirically.

The text chunks for a document are combined with metadata to form a Boolean query. A text search engine (the Lemur Toolkit¹) is used to find candidate near-duplicate efficiently. We call this process *feature-based document retrieval*, because it combines substrings and features extracted during preprocessing (e.g., email senders, receivers, signatures, docket IDs, delivered dates, and email relayers). A query to Lemur looks like: "#AND (docket.oar20020056 router.moveon #OR("standards proposed by" "will harm thousands" "unborn children for" "coal plants should" "other cleaner alternative" "by 90 by" "with national standards" "available pollution control"))"

After a set of candidate near-duplicates is retrieved, the similarity of each candidate to the seed document is measured, using KL-divergence. Similar documents are placed into the seed document's cluster. A preliminary evaluation suggested that the algorithm was generally very effective [22].

2.2 Algorithm Refinement

This section describes refinements to the algorithm described above. The refinements include efficient detection of exact duplicates, a modified similarity measure, and a new clustering algorithm.

2.2.1 Exact Duplicate Removal

Exact duplicates are either unmodified copies of form letters or comments that someone accidentally submitted multiple times (a fairly common event). They are usually a large proportion of most large public comment datasets. The first refinement is to identify exact duplicates very efficiently.

To identity exact duplicates, all white space is removed from the document, and all words in the document are converted into a long string of characters (a *document string*). A hash function is applied to the document string to create a (nearly) unique identifier for this particular document. This process is applied to all documents in the collection. All documents resulting in the same hash value are considered exact duplicates.

The hash function is the security hash function, SHA1 [16], suggested by NIST. It makes sure that the chance of hash value collision is very low and the whole process is secure. It is also designed to be very fast and is good for messages of any length. It is designed for text processing and is known for its even distribution of hash values. SHA1 produces a 20-byte (160-bit) hash value. By using a secure digest algorithm, it reduces the probability of two different token streams creating the same hash value to $p(2^{160})$.

After hashing, there is a <hash-value, document id> tuple for each document. Tuples are sorted by their hash values, then a

¹ <http://www-2.cs.cmu.edu/~lemur/>

simple linear scan of the list is sufficient to identify documents that have identical hash values.

Given a set of exact duplicates, an arbitrary choice determines which one to consider the reference copy – the seed document. Our system selects the document with the earliest timestamp to be the reference copy, annotates it with the number of exact duplicates, and retains it as a candidate for further study. The rest are marked as exact duplicates and are eliminated from further consideration, although they remain available for reference purposes.

2.2.2 Document Similarity Measure

After a set of candidate near-duplicates is retrieved (Section 2.1), the similarity of each candidate to the cluster seed is measured. Our earlier work [22] used a modified version of KL divergence (relative entropy) to measure the similarity of near-duplicate documents. It first selects a seed document, and then measures the similarity between it and every document in its candidate set. For any two documents d_a and d_b with word probability distributions p_a and p_b respectively, the KL divergence measure of the difference between the two probability distributions is:

$$KL(p_a \parallel p_b) = \sum_{w_j \in d_a} p_a(w_j) \log \frac{p_a(w_j)}{p_b(w_j)} \quad (2)$$

Since KL-divergence is non-negative and non-symmetric, [22] defines and uses the minimum value of two KL-divergences as the distance measure between two documents d_a and d_b :

$$dist(d_a, d_b) = \min(KL(p_a \parallel p_b), KL(p_b \parallel p_a)) \quad (3)$$

One flaw in the prior algorithm is that words that appear in p_a but not in p_b are ignored in the calculation. *Block Edits or Key Blocks*, in which a large block of text is added to or deleted from a form letter, are common in this domain, so it is not unusual for a near-duplicate and its reference copy to have unaligned vocabularies. A modification to the similarity measure solves this problem. Instead of assigning zero weights to an unseen word, it is more effective to give the word a probability proportional to its overall probability in a background language model. The KL distance in Equation (2) becomes:

$$\begin{aligned} KL(p_a \parallel p_b) &= \sum_w p_a(w) \log \frac{p_a(w)}{p_b(w)} \\ &= \sum_w p_a(w) \log p_a(w) - \sum_w p_a(w) \log p_b(w) \end{aligned} \quad (4)$$

The first term in Equation 4 depends on document distribution p_a and hence is irrelevant to ranking other documents. It can be dropped and the KL divergence becomes:

$$KL(p_a \parallel p_b) \propto - \sum_w p_a(w) \log p_b(w) \quad (5)$$

$$\text{where } p_b(w) = \begin{cases} p_s(w|d_b) & \text{if } w \text{ is seen} \\ \alpha_d p(w|C) & \text{otherwise} \end{cases} \quad (6)$$

α_d is a coefficient for each unseen word's probability (and also insures that all of the probabilities sum to one). Hence the KL divergence becomes:

$$- \sum_{w \in d_a} p_a(w) \log \frac{p_s(w|d_b)}{\alpha_d p(w|C)} + \log \alpha_d \quad (7)$$

By Dirichlet prior smoothing [1], we have:

$$p_s(w|d_b) = \frac{tf(w, d_b) + \mu p(w|C)}{\mu + |d_b|}, \quad (8)$$

$$\alpha_d = \frac{\mu}{\mu + |d_b|}, \quad (9)$$

$p_a(w)$ and $p(w|C)$ are estimated by maximum likelihood and given by

$$p_a(w) = p(w_j | d_a) = \frac{tf(w_j, d_a)}{\sum_{w_i \in d_a} tf(w_i, d_a)} \quad (10)$$

$$p(w|C) = \frac{\sum_{d_k \in C} tf(w, d_k)}{\sum_{d_j \in C} \sum_{w_i \in d_j} tf(w_i, d_j)} \quad (11)$$

μ is a parameter in Dirichlet smoothing and is set to 1 in this work.

2.2.3 Incorporating Instance-level Constraints

Near-duplicate detection proceeds more smoothly and efficiently when there are clues about which documents are duplicates. In some duplicate-detection scenarios, files that have identical metadata, such as size, date, and base filename are likely to be copies kept on different directories or on different servers.

Clustering algorithms seek to automatically discover underlying patterns in a dataset. Usually, a search is conducted through the space of possible organizations of the data, preferring those that group similar instances and keep dissimilar instances apart. If additional knowledge about the clustering is known beforehand, the clustering algorithm could be more effective and efficient since we can have pruning at the earlier stage. For example, a user may indicate that a certain pair of documents in the dataset is judged to be similar and a certain other pair of documents is judged to arise from separate clusters. Techniques for introducing additional knowledge to perform constrained clustering have primarily focused on formulating and expressing the knowledge by instance-level constraints [23]. As described in [23] these constraints typically take the form of relations such as *must-link* and *cannot-link* that are enforced between pairs of instances.

In a clustering approach to the problem of near duplicate detection, knowledge about the collection characteristics and document attributes can be used to compute instance-level constraints indicating that certain pairs of documents either must be, or cannot be in the same duplicate cluster.

The must-link conditions include the complete containment of the seed document (*key block*), and minor change < 5% word coverage (*minor change*). The cannot-link condition is only includes for documents that have different email relayers or

- A) Initialize the Duplicate Cluster Collection N : $N \leftarrow \emptyset$ and document collection B .
- B) Get the initial seed documents and pick one seed d_i . Note that the non-seed document will be examined in this process and if there is no cluster for it in the first pass, it will be used as seed in the next pass.
- C) Retrieve candidate set S_i for seed document d_i . For each document $s_{ij} \in S_i$,
- a) if $(s_{ij}, d_i) \in Must$,
add s_{ij} into duplicate cluster n_{dk} : $n_{dk} \leftarrow n_{dk} \cup \{s_{ij}\}$
 - b) \forall cluster centroid d_k in N , if $(s_{ij}, d_k) \in Must$
add s_{ij} into duplicate cluster n_{dk} : $n_{dk} \leftarrow n_{dk} \cup \{s_{ij}\}$
 - c) if $dist(s_{ij}, d_i) < \theta_i$
 - \forall cluster centroid d_k in N ,
if $dist(s_{ij}, d_i) > dist(s_{ij}, d_k)$,
add s_{ij} into duplicate cluster n_{dk} : $n_{dk} \leftarrow n_{dk} \cup \{s_{ij}\}$
unless $(s_{ij}, d_k) \in Cannot$
 - else if $dist(s_{ij}, d_i) \leq \min_k(dist(s_{ij}, d_k))$ and $d_i \notin N$,
create a new cluster n_{di} , add it into N : $N \leftarrow N \cup \{n_{di}\}$,
add s_{ij} into n_{di} : $n_{di} \leftarrow n_{di} \cup \{s_{ij}\}$ unless $\exists d_l \in n_{di} (s_{ij}, d_l) \in Must$
 - eliminate s_{ij} from n_{dk} : $n_{dk} \leftarrow n_{dk} - \{s_{ij}\}$, unless $\exists d_l \in n_{dk} (s_{ij}, d_l) \in Cannot$
- E) If $B = \emptyset$, output N as the final set of duplicate clusters.

Figure 1: Algorithm to form duplicate clusters

docket ids. Note that instance-level constrained clustering is very flexible; as more background knowledge about the dataset is acquired, it can be added as new constraints.

3. EVALUATION METHODOLOGY

Clustering and near-duplicate detection algorithms are difficult to evaluate because “ground truth” information is rarely available. This section describes the datasets used in our experiments, our evaluation metrics, and in particular our methodology for acquiring “ground truth” assessments.

3.1 Data Sets

Our research was conducted with a public comment dataset for the U.S. Environmental Protection Agency’s (EPA) proposed National Emission Standards For Hazardous Air Pollutants For Utility Air Toxics rule (USEPA-OAR-2002-0056, “Mercury rule”). The dataset contains 536,975 email messages. The algorithm successfully ran on the entire dataset.

However, it is impractical to have human assessment on the entire dataset. To make the human assessment doable, two random samples of size 1,000 were generated as evaluation set. The exact duplicates were also removed and left 275 and 270 documents in what became known as the *NTF* (*Name That Form*) and *NTF2* subsets. Table 1 provides statistics about these datasets. Section 3.3 provides additional description of these datasets.

3.2 Evaluation Metrics

The accuracy of near-duplicate detection was measured using well-known evaluation metrics such as Precision, Recall and F1-

Table 1: Sample Dataset Statistics

Sample Set Name	NTF	NTF2
Source	USEPA-OAR-2002-0056, “Mercury rule”	USEPA-OAR-2002-0056, “Mercury rule”
# of documents originally	1000	1000
# of documents after exact duplicates removal	275	270
# of reference copies (document with > 5 exact duplicates)	28	26
average document length before removing header/signature lines	220	213
average document length after removing header/signature lines	156	152
# unique terms	3330	3437

measure [22], and intercoder agreement metrics such as Cohen’s Kappa and AC1. The experiments were designed to determine whether the system’s agreement with human assessors was comparable to agreement between two or more human assessors.

Cohen’s Kappa: Cohen’s kappa statistic [4] is often used to evaluate clustering effectiveness. It assesses agreement between two sets of results or in another word, two coders. Therefore in our experiments, the intercoder agreements are always measured between two coders. The kappa coefficient is defined as:

$$\kappa = \frac{p(A) - p(E)}{1 - p(E)} \quad (12)$$

where $p(A)$ is the observed agreement between the two assessments, a is the number of pairs in the same group in the ground truth and in the clustering (agreement), b is the number of pairs in the same group in the ground truth but different in the clustering (false negative), c is the number of pairs in the different groups in the ground truth but the same in the clustering (false positive), d is the number of pairs in the different groups in the ground truth and in the clustering (agreement). $p(A)$ can be calculated as $(a+d)/m$ and $m=a+b+c+d$. $p(E)$ is the agreement expected by chance, and is calculated as:

$$p(E) = (a+b)(a+c)/m^2 + (b+c)(c+d)/m^2 \quad (13)$$

AC1: Cohen’s kappa suffers from problems of bias and prevalence. If the agreement between assessors is high but skewed to a few categories, as is common in public comment datasets, the resulting kappa value cannot truly represent the degree of agreement [9]. AC1 corrects this problem and calculates the chance agreement in another way:

$$p(E) = 2P1(1-P1) \text{ where } P1 = ((a+b)+(a+c))/2m. \quad (14)$$

Although kappa is not appropriate for our task, it is used often in prior research, so we report both kappa and AC1 values. We

P59: 029932.txt - 59:1 [The misuse of power here is incredible!!!! Doesn't the welfare of our children come first? Enact more regulations for big business and keep our children safe!!!! There is no excuse for such blatant mishandling of this situation.. Act Now!!!!]
Codes: [Disappointment] [Economic] [Public Health & Safety] [Social Values] [Strength=High] [Unique Text in a Form Letter]
No memos
The misuse of power here is incredible!!!! Doesn't the welfare of our children come first? Enact more regulations for big business and keep our children safe!!!! There is no excuse for such blatant mishandling of this situation.. Act Now!!!!

Figure 2: Human Annotation Example

believe that AC1 is a reliable measure for this task, thus we mainly rely on it to draw conclusions.

3.3 Human Annotation Methodology

The authors developed a manual annotation (*coding*) methodology to guide student annotators in identifying near-duplicate public comments and unique texts. The coding system was developed and deployed through an iterative process using coders at the University of Pittsburgh's Qualitative Data Analysis Program (QDAP). The QDAP coders used ATLAS.ti, a commercial off-the-shelf qualitative data analysis application, to capture and report their annotations.

3.3.1 Evolution of the Coding Scheme

In its earliest iteration, during the spring of 2005, the current coding scheme was designed primarily to serve social science goals related to the project. It contained 28 distinct sub-topic and discourse style codes, such as "Legal," "Economic," "Public Health & Safety," "Personal Experience," and "Strength-High," as well as a code for "Unique Text in a Form Letter." Five coders were trained to identify sub-paragraph and paragraph level instances of these codes in a random sample of 1,000 e-mails from a different dataset. The sample was divided to ensure a unique 2-coder overlap on 320 documents. In addition, at the document level, the coders were expected to identify whether the document was an exact duplicate or a near duplicate. An example of this earliest coding scheme is shown in Figure 2.

Analysis of the first round of coding revealed that many of the sub-topic and document-level codes were applied inconsistently. The overall measure of inter-rater reliability was extremely low. This necessitated further training sessions and a more thorough clarification of the coding heuristics. After the retraining of the coders and returning them to review and correct their annotations, the F-measure of inter-rater reliability for all codes combined (including overlapping spans of text) remained low (0.53). While the coders were still struggling to consistently apply some of the amorphous and insufficiently defined subtopic codes, they showed promise identifying stakeholders (0.77) and unique text added to a form letter (0.89). The 28 subtopic codes required multiple passes by the coders on codes and clusters to correct the many errors of the first unwieldy round of coding. Altogether they looked at the initial set of texts 3 times and still produced generally unreliable coding.

For the purpose of preparing this paper, a new coding approach was developed in mid-August 2005. Using lessons from the first

round, a second coding was specifically tailored to the needs of researchers developing and evaluating near-duplicate detection tools and federal agencies trying to manage large public comment campaigns. In the modified coding scheme, a new sample of 1,000 e-mails was selected and the exact duplicates were removed using the exact-duplicate detection algorithm described above (Section 2.2.1), leaving 275 documents in the NTF pool.

In the first NTF round, coders were provided with a set of 28 "known form letters", defined as comments that had 5 or more exact duplicates in the dataset. Two coders were trained to apply a single code (Unique Text). The NTF experiment also required the coders to associate each comment with one of the 28 known form letters, or to identify it as unique. Three further document-level labels were also used ("Block Edit," "Minor Change," and "Singleton"²). The initial NTF round, with its narrow focus on coding for near-duplicate detection, predictably produced a high measure of inter-rater reliability (0.82) for the unique text code on the first pass, with no need to retrain the coders or redo the coding. It also highlighted the need to better specify a precise rule set that more effectively defined the proper spans of text to be annotated. In the NTF round, one coder annotated almost 50% more text spans than the other. Note that the NTF rounds were completed with a single pass on clusters and a single pass on codes. Different from the previous 28 sub-topic coding scheme, which requires multiple annotation passes to reach a reasonable intercoder agreement, NTF round coding resulted in higher intercoder agreement at the first time.

A review of the NTF coding raised the possibility that introducing a second code ("Signature") might reduce the variance between the coders and produce a more accurate annotation for the code "Unique Text." The NTF2 round did result in high measures of reliability for both "Signature" (0.95) and "Unique Text" (0.87).

By early November, a decision was made to repeat the coding of the NTF and NTF2 samples using a new set of six coders who had no prior NTF experience. Three of the six were new to coding in QDAP, having just completed their ATLAS.ti training, while the remaining three were relatively experienced and acted as mentors for the new coders. A more precisely specified rule set was generated and relayed to the experienced coders. For both the new NTF and NTF2 rounds, two and four coders were used respectively. Each of these experiments required coders to use a total of four codes, with the addition of "Header" and "Stakeholder" to "Unique Text" and "Signature." Both rounds offered unique insights into the techniques for training coders to produce reliable annotations. Moreover, as mentioned in the previous sections, to group duplicates into subcategories by the editing styles is a common strategy. In the two datasets NTF and NTF2, human assessors identified the block added, block deleted, minor change, block rearrange and singleton document and "repeated copies". Note that "repeated copies" refers to a special situation where the form letters are repeated several times in the submitted public comments and the public comments contain only the repeated paragraphs.

² We call it *singleton* since it forms a cluster by itself alone in the clustering algorithm.

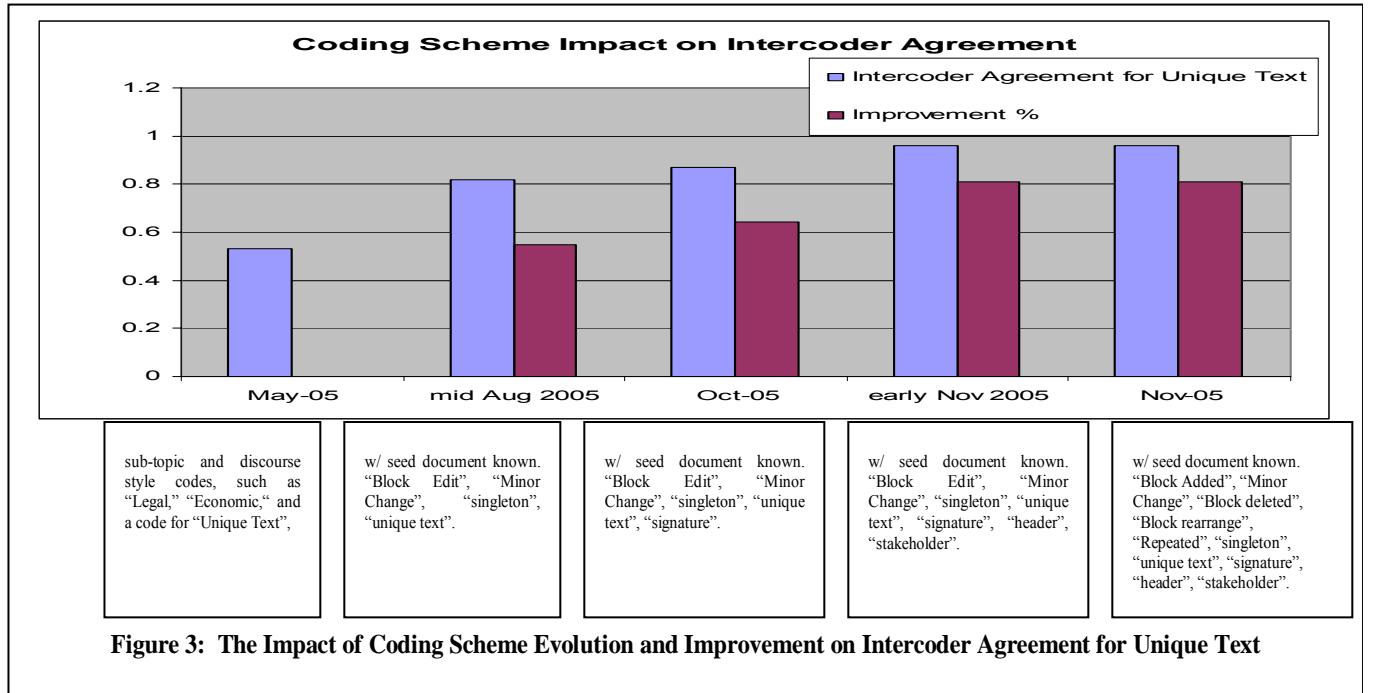


Figure 3 illustrates the impact of the coding scheme evolution and the improvement gained in the intercoder agreement, in particular, for the code of “unique text” only. The more precise coding scheme produces much higher and more intercoder agreements between human annotators.

The latest NTF round produced the most reliable coding to date of any pair of QDAP coders working on any project. As reported in Table 2 the F-measure for exact matches (0.91) and matches including overlapping text spans (0.98), suggest this is a gold standard for what two coders with a consistent, narrow and very precise rule set can accomplish. The NTF2 round (Table 3) also resulted in solid inter-rater-reliability coefficients, with an average F-measure of 0.90 across all four codes and coders.

3.3.2 Training Effects

Repeated rounds of coding for near-duplicates taught us something important about training effects in manual annotation. In new NTF2, experienced coders UCSUR 8 & 9 received

instructions directly from the QDAP Director and then independently relayed them to UCSUR 15 & 17. This to say, UCSUR 8 trained 17 and 9 trained 15.

As Table 3 shows, there is a pattern perhaps reflecting differences in how 8 trained 17 and 9 trained 15 to code “Unique Text” since the number of “unique text” found by coder 8 is close to 17’s while 9’s is closer to 15’s. It is interesting to study more the training effects among the coders. We conducted 4 intercoder agreements for each pair of the coders. They are: kappa (w/o overlap), which is the cohen’s kappa agreement of two coders for the texts that exact match; kappa, which is cohen’s kappa for the texts including partial match (a more Lenin evaluation), F-measure (w/o overlap), which is the F-measure of two coders for the texts that exact match by using one coder as the ground truth and calculate the other one using traditional F-measure in information retrieval; F-measure, which is F-measure for the texts including partial.

Table 2: NTF Coding Results

Code	# found by 13	# found by 16	Kappa (w/o overlap)	Kappa (w/ overlap)	F-measure (w/o overlap)	F-measure (w/overlap)
Header	266	266	0.93	0.99	0.94	0.99
Signature	281	283	0.94	0.98	0.95	0.98
Unique Text	218	220	0.86	0.96	0.86	0.96
Total	798	799	0.92	0.98	0.91	0.98

Table 3: NTF2 Coding Results with Training Effects

Code	# found by 8	# found by 9	# found by 15	# found by 17	Avg F-measure (w/o overlap)	Avg F-measure (w/overlap)
Header	264	262	264	265	0.93	0.99
Signature	276	281	275	277	0.91	0.96
Unique Text	213	146	171	214	0.61	0.72
Total	781	715	723	778	0.83	0.90

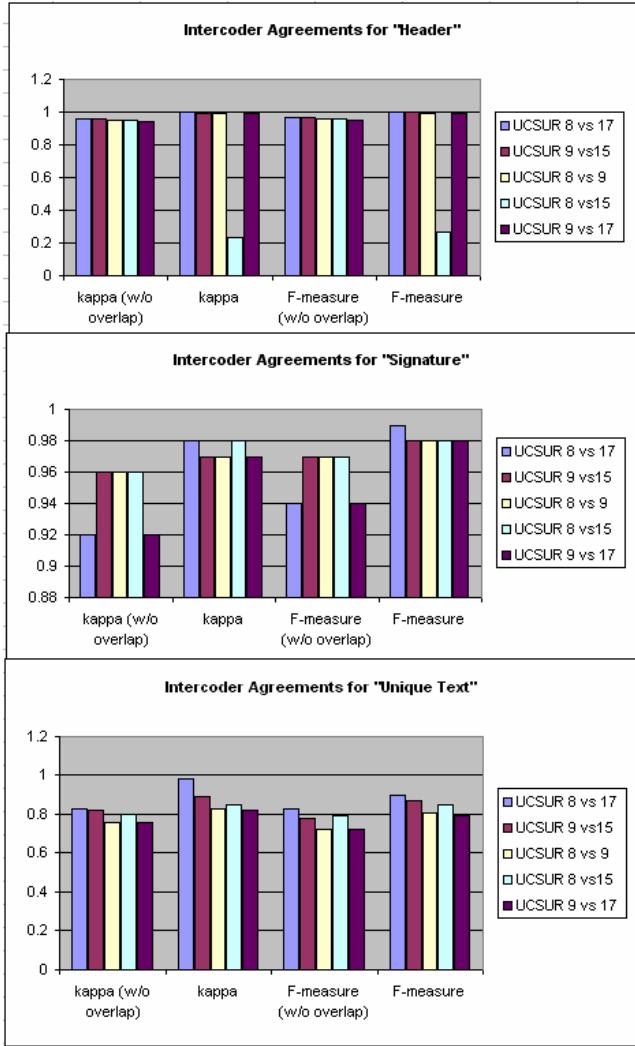


Figure 4: Training Effects on Intercoder Agreements

Figure 4 illustrates the possible presence of training effects on the coding of “header”, “signature” and “unique text”. We can see the most obvious effects on the “unique text”, where the pairings of coder 8 vs. 17 and coder 9 vs. 15 give the top two intercoder agreements among all four kinds of agreements. Interesting enough to notice that the intercoder agreements of coder 8 and 15 are also higher than that of the two senior coders (8 and 9) who are directly trained by the QDAP Director. We are not sure whether it means that coder 15 actually understands the coding scheme best. However, we do use the annotation from coder 15 as a gold standard for the later duplicate detection tests.

3.3.3 Duplicate Editing Styles

In the latest runs, the subcategories of near-duplicates, or in other words, the editing styles of near-duplicates are considered as part of the coding scheme. It is interesting to see what makes up the near-duplicates. It is also very important for us to study the effectiveness of the automatic near-duplicate detection algorithm on each subcategory. Figure 5 and Figure 6 show the duplicate editing style distribution for the NTF and NTF2 datasets. The

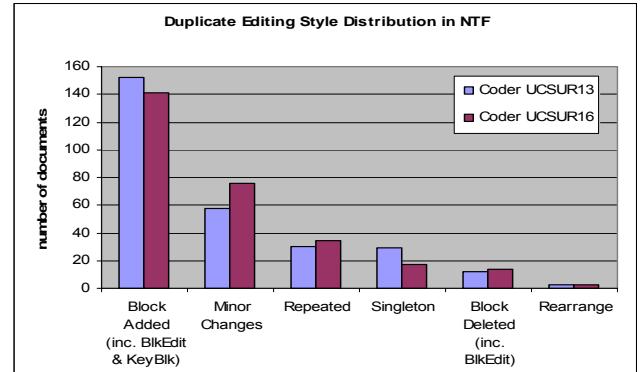


Figure 5: Duplicate Editing Style Distribution in NTF

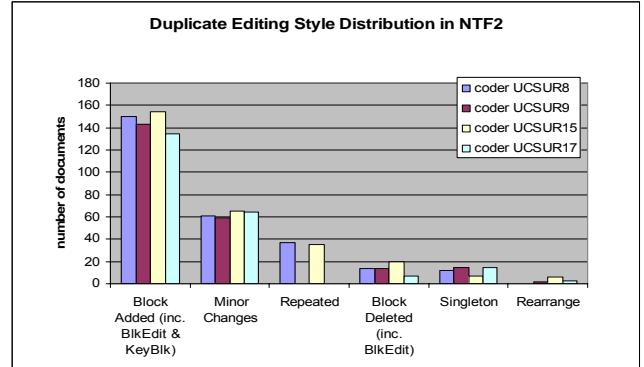


Figure 6: Duplicate Editing Style Distribution in NTF2

annotations from coders UCSUR13 and UCSUR16 are used for NTF and annotations from coders UCSUR8, UCSUR9, UCSUR15 and UCSUR17 are used for NTF2. Based on the two figures, we have the following observations.

- Block Added is the dominant editing style for people submitting public comments. Block Edit and Key Block both belong to this category.
- Minor Change is the next major editing style for near-duplicates in the public comment domain.
- Singleton documents are unique comments not based on a form letter. They could contain unique insights and opinions, derivative opinions, spam or viruses. The amount of singleton documents in both datasets is reasonably large. These are of potentially of interest for social science research. Another important source of possibly-unique opinions added to form letters (Block Added).
- Note that for NTF2, two coders, UCSUR9 and UCSUR17 considered that no documents belong to the “exact copy” category while the other two coders considered that there are at least 30+ documents in this category. This is an interesting disagreement that we cannot yet explain.

4. EXPERIMENTAL RESULTS

4.1 System-to-Human Intercoder Agreement

The first set of experiments explored the agreement between DURIAN and human assessors at identifying near-duplicate comments, unique passages in near-duplicate comments, and unique comments. In these experiments DURIAN was considered another coder, and accuracy is measured using the intercoder agreement metrics discussed previously. Experiments were conducted on the NTF and NTF2 datasets. NTF has 28 form letters while NTF2 has 26.

Usually intercoder agreement is calculated across all pairs of documents (*micro-averaging*). However, it is also useful to measure intercoder agreement for each cluster and then to average over the number of clusters (*macro-averaging*). Macro-averaging focuses on agreement on the large letter-writing campaigns, whereas micro-averaging gives a better measurement over the set of letter-writing campaigns.

In the following tables, we use Coder A to represent UCSUR 13 for NTF and UCSUR8 for NTF2; and Coder B to represent UCSUR 16 for NTF and UCSUR9 for NTF2. We pair them this way because these are the best combinations of assessors in terms of intercoder agreement between humans. They are the gold standard.

Table 4: Duplicate Detection Intercoder Agreement

	Averaged across all clusters				Averaged across all pairs			
	Kappa(Cohen)		AC1		Kappa (Cohen)		AC1	
	NTF	NTF2	NTF	NTF2	NTF	NTF2	NTF	NTF2
Coder A/ Coder B	0.53	0.19	0.93	0.90	0.99	0.97	0.99	0.95
Coder A/Program	0.52	0.15	0.92	0.80	0.90	0.89	0.93	0.90
Coder B/Program	0.53	0.17	0.90	0.82	0.92	0.90	0.91	0.91

Table 4 summarizes the first set of experimental results. The macro-averaged Cohen’s kappa values are surprisingly low for both the human-to-human and program-to-human intercoder agreements. This is surprising because the agreement between coder A and B about whether a document belongs to a certain cluster is actually high in raw numbers, and there is prevalence in the agreement categories. The agreement about the number of documents in the same cluster is much higher than the agreement about the number of document not in the same cluster (i.e., $a >> d$ in the kappa value calculation). This unbalanced distribution causes Cohen’s kappa to fail to represent the true agreement.

The macro-averaged AC1 values are high. Given that AC1 is a stable agreement measure even when prevalence and bias problems are present, we believe that this metric accurately reflects the degree of agreement between humans and DURIAN.

In the pairwise comparison, Cohen’s kappa and AC1 both show high intercoder agreements. This is very desirable, however the result might be misleading. There are some very large letter writing campaigns in these datasets, which creates huge duplicate clusters. Huge clusters tend to dominate the final result of pairwise comparison since they have more pairs. Thus macro-averaged AC1 is probably the best measure to use for evaluating

the near-duplicate clustering. DURIAN’s agreement with human coders as about good as the agreement between pairs of humans.

In addition to identifying near-duplicates and forming duplicate clusters, DURIAN is able to recognize header blocks, signature lines, and unique text in a public comments. We again use both Cohen’s kappa and AC1 as the evaluation metrics. Tables 5-7 summarize the results.

Table 5: Unique Text Intercoder Agreement

	Kappa (Cohen)		AC1	
	NTF	NTF2	NTF	NTF2
Coder A/ Coder B	0.96	0.83	0.98	0.82
Coder A/Program	0.80	0.76	0.86	0.74
Coder B/Program	0.78	0.75	0.84	0.74

Table 6: Header Detection Intercoder Agreement

	Kappa (Cohen)		AC1	
	NTF	NTF2	NTF	NTF2
Coder A/ Coder B	0.99	0.99	0.99	0.99
Coder A/Program	0.93	0.92	0.93	0.91
Coder B/Program	0.93	0.92	0.93	0.91

Table 7: Signature Line Detection Intercoder Agreement

	Kappa (Cohen)		AC1	
	NTF	NTF2	NTF	NTF2
Coder A/ Coder B	0.98	0.97	0.99	0.97
Coder A/Program	0.90	0.92	0.90	0.91
Coder B/Program	0.90	0.90	0.90	0.89

Unique text is the text a person added to a form letter; it might raise a substantive issue, so it requires agency review. The human assessors have a high agreement in NTF and a lower but still very good agreement on NTF2. However, the agreements of the program with both human assessors are lower, perhaps just above what we would consider acceptable. The human assessors only consider a major addition of text to the original form letter to be “unique text”. The program, however, is sensitive to changes varying from the large block changes to even small punctuation changes. Table 5 suggests that DURIAN might benefit from some tuning to be less sensitive to small changes that human coders consider trivial.

As a preprocessing step, DURIAN’s header and signature detection algorithms were very effective. Although pleasing, this result may simply indicate that that public comments based on form letters mainly follow a formal letter-writing style that makes it easy to detect the header and signatures lines.

Cohen’s Kappa and AC1 tend to agree about unique text, header and signature line detection accuracy. Pairwise comparisons were used for Table 5-7 because cluster-based comparisons are less meaningful when evaluating header, signature, and unique text capabilities. The skewed distribution problem does not happen in this case, thus Cohen’s kappa is reliable.

4.2 Duplicate Detection Algorithms

This paper presents DURIAN, a new near-duplicate detection algorithm designed for public comment datasets and notice and comment rulemaking tasks. However, other algorithms also detect duplicate documents and might be applied to this task. A set of experiments was conducted to evaluate several well-known duplicate-detection algorithms on the NTF and NTF2 datasets. The contenders are described below.

Full fingerprinting (full): Every substring of size s in the documents are selected and hashed. s is set to 3 in our experiments. Every hash value (a fingerprint) is stored for the document in a form of <fingerprint, document id> tuple. Every substring will contributes one such tuple since every substring resulting one fingerprint. Therefore we have a huge list. The duplicate detection is performed by 1) sorting the tuples <fingerprint, document id>; 2) generating overlapping fingerprint records, if a document with id =567 contains a fingerprint in the form letter, a tuple <567, 1> is generated; 3) counting the overlap fingerprint records for all documents, we get <document id, count>. If the count of the overlap fingerprints in a document to the form letter is above 80%, it is considered as a duplicate.

Shingling (DSC) [2]: Every 5 overlapping substring of size s in the documents is selected and hashed. s is set to 3 again in our experiments. The hash values are stored for each document. Duplicate detection is performed in the same way as described in full fingerprinting. If the count of the overlap fingerprints in a document to form letter is above 80%, it is considered as a duplicate to form letter.

I-Match [3]: The N words with the highest idf values in a document are selected, N is set to 30 in our experiments. Note that the top 5 idf words are ignored here since they might be some random mistakes such as misspellings. A single fingerprint is generated for each document. Duplicate detection is performed by sorting all <fingerprint, document id> tuples. Those documents agree with the fingerprint of the form letter, are selected as the (near) duplicates.

Durian: The algorithm proposed in this paper.

The parameters used in all the experiments for competing methods were tuned using parameter sweeps and/or the best values reported in other researcher's work [10][6].

In this experiment we assume that the form letters are known, and that the task is to identify the near-duplicates identified by the human coders UCSUR16 (NTF) and UCSUR15's (NTF2). In order to study the effectiveness of duplicate detection techniques on different duplicate categories, the detection results are further distributed into different duplicate categories. The average precision, average recall, and average F1-measure for each category averaged for each form letter are reported here.

Full fingerprinting was the most simple substring selection technique. It gave the largest possible set of fingerprints for a document. Not surprisingly, it either gives the best or the second best F1 value in every category. Full finger printing is very effective, however it is also the most computationally expensive method. Since every substring is stored as a hash number, the effort of sorting a huge list of tuples is unavoidable. Both the storage and execution time is very costly.

Durian consistently performs well on all the categories, occasionally beating the full fingerprint approach. However, the retrieval and detection time is much lower than for full fingerprinting.

Two other techniques DSC and I-Match were not as effective as full fingerprinting and DURIAN. In general, DSC outperforms I-Match. I-Match is very sensitive to both Block Added and Block Deleted. It is also very sensitive to Text Minor Change. When the changed words are critical, i.e., appear in the fingerprint that I-Match selected, the algorithm fails to detect the near-duplicates. In general, I-Match produced fairly low Precision and the Recall.

Table 8: Comparison of duplicate detection technologies

Duplicate category	Algorithm	Avg Precision		Avg Recall		Avg F1	
		NTF	NTF2	NTF	NTF2	NTF	NTF2
Exact	Full	1	1	1	1	1	1
	DSC	0.97	0.98	0.98	0.98	0.97	0.98
	I-Match	0.91	0.9	0.8	0.75	0.85	0.82
	DURIAN	1	1	1	1	1	1
Minor change	Full	0.95	0.95	0.95	0.95	0.95	0.95
	DSC	0.9	0.9	0.9	0.9	0.9	0.9
	I-Match	0.79	0.8	0.76	0.78	0.77	0.79
	DURIAN	0.95	0.95	1	1	0.98	0.97
Block Added	Full	0.97	0.98	0.98	0.98	0.98	0.98
	DSC	0.73	0.7	0.74	0.78	0.73	0.74
	I-Match	0.32	0.35	0.4	0.42	0.36	0.38
	DURIAN	0.98	0.98	0.98	0.98	0.98	0.98
Block Deleted	Full	0.9	0.9	0.9	1	0.98	0.95
	DSC	0.72	0.75	0.74	0.78	0.73	0.76
	I-Match	0.3	0.33	0.4	0.4	0.36	0.36
	DURIAN	0.98	0.98	0.98	0.98	0.98	0.98
Singleton	Full	0.9	0.9	0.9	0.9	0.93	0.9
	DSC	0.72	0.74	0.8	0.8	0.76	0.77
	I-Match	0.84	0.88	0.78	0.8	0.81	0.84
	DURIAN	0.94	0.94	0.94	0.94	0.94	0.94
Rearrange	Full	1	1	1	1	1	1
	DSC	0.67	0.83	0.67	0.83	0.67	0.83
	I-Match	1	1	1	1	1	1
	DURIAN	1	1	1	1	1	1

5. CONCLUSION AND NEXT STEPS

U.S. regulatory agencies are required to solicit, consider, and respond to public comments before issuing regulations. Recently, the shift from paper to electronic public comments makes it much easier for individuals to customize form letters while harder for agencies to identify substantive information since there are many near-duplicate comments that express the same viewpoint in slightly different language.

This research focused on the process of identifying near-duplicates of form letters, and unique passages added to form letters. An extensive study of human intercoder agreement on public comments provided by the Environmental Protection Agency set a baseline against which to evaluate automated techniques. This paper demonstrates that statistical similarity measures and instance-level constrained clustering can be quite

effective for efficiently identifying near-duplicates. In some tasks the algorithm is comparable to our best human assessors; in other tasks it is slightly less effective than our best assessors, but perhaps sufficiently effective for production use. When the algorithm “fails”, it tends to do so by classifying a modified form letter as a unique comment, thus referring it for human review.

The current study reports results against two samples of one corpus of public comments to the EPA. We are currently at work on evaluation using another corpus provided by another agency. Additional experiments will provide greater insight into the strengths, weaknesses, and generality of the algorithm.

Whether, how or when this type of technology will emerge as a factor in regulatory rulemaking is beyond the scope of this paper. Optimists may read this report as a bellwether signaling the imminent end of a temporarily vexing problem. Mass comment campaigns will be more manageable when tools such as ours make a successful technology transfer into the hands of agencies receiving the comments. Furthermore, if agency personnel are so inclined, the addition of unique stakeholder views to mass comment campaigns will come into much greater focus with much less effort and expense.

ACKNOWLEDGMENTS

We are grateful to the USDA, US DOT, and US EPA for providing the public comment data that made this research possible. This research was supported by NSF grants EIA-0327979 and IIS-0429102. Any opinions, findings, conclusions, or recommendations expressed in this paper are the authors', and do not necessarily reflect those of the sponsor. We are also grateful to invaluable comments from the anonymous reviewers.

6. REFERENCES

- [1] S. Brin, J. Davis, and H. Garcia-Molina. Copy detection mechanisms for digital documents. In Proceedings of the Special Interest Group on Management of Data (SIGMOD 1995), pages 398–409. ACM Press, May 1995.
- [2] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig. Syntactic clustering of the web. In Proceedings of WWW6 '97, pages 391–404. Elsevier Science, April 1997.
- [3] A. Chowdhury, O. Frieder, D. Grossman, and M. McCabe. Collection statistics for fast Duplicate document detection. In ACM Transactions on Information Systems (TOIS), Volume 20, Issue 2, 2002.
- [4] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46, 1960.
- [5] C. Coglianese, E-Rulemaking: Information Technology and the Regulatory Process. *Administrative Law Review* 56(2): 353-402. 2004.
- [6] J. Conrad and C. P. Schriber. Online duplicate document detection: signature reliability in a dynamic retrieval environment. Proceedings of the twelfth international conference on Information and knowledge management, Pages: 443 - 452 New Orleans, LA, USA, 2003.
- [7] F. Emery and A. Emery, A Modest Proposal: Improve E-Rulemaking by Improving Comments. *Administrative and Regulatory Law News*, 31(1): 8-9. 2005.
- [8] Government Accountability Office. Electronic Rulemaking: Progress Made in Developing a Centralized E-Rulemaking System. GAO-05-777, 2005.
- [9] K. Gwet. Kappa Statistic is not Satisfactory for Assessing the Extent of Agreement between Raters. *Statistical Methods for Inter-rater Reliability Assessment*, No.1, April 2002.
- [10] T. Hoad and J. Zobel. Methods for identifying versioned and plagiarized documents. In *Journal of the American Society of Information Science and Technology*, Volume 54, Issue 3, 2003.
- [11] C.M. Kerwin, Rulemaking: How Government Agencies Write Law and Make Policy 3rd Ed. CQ Press, Washington, DC, 2003.
- [12] G.T. Lau, K.H. Law, and G. Wiederhold,. A Relatedness Analysis Tool for Comparing Drafted Regulations and Associated Public Comments. *I/S* 1(1): 95-110. 2005.
- [13] J.S. Lubbers, A Guide to Federal Agency Rulemaking. Third Edition. Chicago, ABA, 1998.
- [14] U. Manber. Finding similar files in a large file system. In 1994 Winter USENIX Technical Conference, pages 1-10, San Francisco, CA, January 1994.
- [15] D. Metzler, Y. Bernstein and W. Bruce Croft. Similarity Measures for Tracking Information Flow, Proceedings of the fourteenth international conference on Information and knowledge management, CIKM'05, October 31.November 5, 2005, Bremen, Germany.
- [16] NIST, “Secure Hash Standard”, Federal Information Processing Standards Publication 180-1, 1995.
- [17] N. Shivakumar and H. Garcia-Molina. SCAM: a copy detection mechanism for digital documents. In Proc. International Conference on Theory and Practice of Digital Libraries, Austin, Texas, June 1995.
- [18] S. Shulman, L. Thrane, and M.C. Shelley. eRulemaking, in G. David Garson (Ed.) *The Handbook of Public Information Systems* 2nd Ed. CRC Press, Boca Raton, FL, 2005, 237-254.
- [19] S.W. Shulman, E-Rulemaking: Issues in Current Research and Practice. *International Journal of Public Administration* 28: 621-641. 2005.
- [20] S.W. Shulman, The Internet Still Might (But Probably Won't) Change Everything. *I/S* 1(1): 111-145. 2005.
- [21] S.W. Shulman, An Experiment in Digital Government at the U.S. National Organic Program. *Agriculture and Human Values* 20(3): 253-265, 2003.
- [22] H. Yang and J. Callan. Near-Duplicate Detection for eRulemaking. In Proceedings of the 5th National Conference on Digital Government Research (DG.O2005), Atlanta, GA, USA, 15-18 May 2005.
- [23] K. Wagstaff and C. Cardie, 2000. Clustering with instance-level constraints. In Proceedings of ICML-2000. pp. 1103–1110. Palo Alto, CA.
- [24] C. Zhai and Lafferty, J. (2001b). A study of smoothing methods for language models applied to ad hoc information retrieval. In Proceedings of SIGIR 2001, pages 334-342.

Progress in Language Processing Technology for Electronic Rulemaking

Stuart Shulman
University of Pittsburgh
121 University Place, Suite 600
Pittsburgh, PA 15260
shulman@pitt.edu

Jamie Callan
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213-8213
callan@cmu.edu

Eduard Hovy
USC-ISI
4676 Admiralty Way
Marina del Rey, CA 90292-6695
hovy@isi.edu

Stephen Zavestoski
University of San Francisco
2130 Fulton Street
San Francisco, CA 94117-1080
smzavestoski@usfca.edu

ABSTRACT

In this project, we are developing new text processing tools that help people perform advanced analysis of large collections of text commentary. This problem is increasingly faced by the U.S. federal government's regulation writers who formulate the rules and regulations that define the details of laws enacted by Congress. Our research focuses on text clustering, text searching, near-duplicate detection, opinion identification, stakeholder characterization, and extractive summarization, as well as the impact of such tools on the process of rulemaking itself. Versions of a Rule-Writer's Workbench are being built by researchers at ISI and CMU, made available for experimental use by our government partners at the DOT and EPA, and evaluated by researchers at the Library and Information Science and Sociology departments at the universities of Pittsburgh and San Francisco, respectively. This project started in October 2004 and is funded for 3 years under the NSF's Digital Government Program.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Content Analysis and Indexing – *indexing methods, linguistic processing*.

General Terms

Algorithms, Experimentation.

Keywords

Information retrieval, information extraction, duplicate detection, opinion recognition, regulatory rulemaking, federal government.

1. MOTIVATION AND BACKGROUND

Many people today—including news analysts, opinion pollsters, advertisers, and government regulation writers—need to interpret, structure, and rapidly master large quantities of commentary about their work. We focus on the task facing several thousand regulation writers who formulate the rules and regulations required for the implementation of laws.

In this procedure, they invite and then process in detail comments from the public on their proposed regulations. In some instances, they may receive several hundred thousand form letters by email or studies of a few hundred or thousand pages.

A decade ago form letters were not difficult to process. Except for the signature, address and possibly a brief hand-written comment, they were exact duplicates of the original form letter. Form letters could be identified and sorted easily, so that the original could be considered, and the duplicative copies counted, reported, and then largely ignored.

Since our last report, there continue to be electronic mass comment campaigns. For example, the USDOT has a new Corporate Average Fuel Economy (CAFE) rule forthcoming that received over 40,000 public comments. As part of its continuing support for this research, the DOT expeditiously provided our group with this data for future experiments and hosting on the eRulemaking Testbed at CMU.

We have also been in a discussion with personnel at the Union of Concerned Scientists about its use of web-based form letter campaigns. This conversation seeks to advance the activist-business-government-researcher dialogue generated as by-product of the past 5 years of NSF-funded studies of electronic rulemaking.

This particular multi-year study will expand and upgrade the scope and function of CMU's eRulemaking Testbed and systematically feedback information from users in and out of government via workshops, focus groups, Web surveys, reports, publications, and presentations.

Our research explores the use of information extraction and information retrieval to develop tools that assist rule-writers and analysts in managing large volumes of public comments. Information extraction techniques strip off email headers, salutations, signature lines, and advertising text. Text clustering algorithms identify exact duplicates, group together comments that are similar but not identical, and organize them hierarchically for browsing by rule-writers. Text-differencing algorithms identify where a person has edited a form letter so that a rule-writer's attention is drawn immediately to the unique part of an edited form letter.

Our project aims to develop tools for interpreting, structuring, and rapidly mastering large quantities of opinion-based text. This study builds on our previously developed eRulemaking testbed by systematically collecting information from users of the testbed's text analysis tools. The testbed's new text processing tools perform text clustering, text searching using information retrieval, near-duplicate detection, opinion identification, stakeholder characterization, and extractive summarization of large volumes of public commentary.

2. WORK TO DATE

The social science component of the project convened a focus group in September 2005. The subjects were citizen commenters from an earlier Union of Concerned Scientists mass comment campaign who were chosen at random from across the U.S. The objective was to understand more thoroughly the motivations of mass campaign commenters, their decisions in using and modifying form letters, and the expectations they have regarding the way in which agencies respond to their comments.

A new Qualitative Data Analysis Program (QDAP), initiated by Dr. Stuart Shulman at the University of Pittsburgh, continues large scale manual coding of thousands of public comments and press accounts of the rulemakings under study. These manual annotation techniques were further refined in order to improve inter-coder reliability and to increase their utility to natural language processing researchers. We continue to experiment with training and annotation innovations and hope to host a workshop on this specific task in the near future. The challenge has been to develop tailored, reliable coding schemes that can serve both the social and computer science research communities.

Our previous near-duplicate detection work was extended significantly this year, resulting in DURIAN (**D**UPLICATE **R**EMOVAL **I**n **l**ARGE collectioN), an algorithm using a traditional bag-of-words document representation, document attributes ("metadata"), and document content structure to identify form letters and their edited copies in public comment collections. In Yang et al. (under review), we report results against two samples drawn from a large set of public comments submitted to the EPA during the summer of 2005. We plan further evaluation using corpora provided by another agency. Early tests suggest that DURIAN is about as effective as human assessors at detecting duplicates. Further analysis will provide greater insight into the strengths, weaknesses, and generality of the algorithm.

The second substantive accomplishment has been the introduction of the eRuleClient tool, through which QDAP coders can more easily annotate both the subtopic codes and argument linkage structures of public comments. Through several iterations with

QDAP coders implementing the EruleClient tool and researchers and programmers at ISI making on-the-fly updates, coder satisfaction with the tool was greatly enhanced.

Two further accomplishments are the creation of technology that categorizes each sentence into one of nine topic categories (e.g., Economic, Environment, Technology, etc.), and the deployment of opinion detection software, allowing the user to identify all fragments that are positive or negative about the overall theme. This engine achieves an F-score of 71% (Kwon et al., under review). Work underway seeks to determine the argument structure (main topic, subtopics, and dependencies) of the texts.

The next step of the research is to combine all the above work in order to create a system that performs multi-aspect analysis of rulemaking comments and provides a useful review tool for rulemakers, and then to test its efficacy with government users.

3. PRESENTATIONS IN 2005

We have presented this work at the following venues:

American Bar Association's Administrative Law Conference

American Political Science Association

European University Institute

International Conference on E-Gov Research

Judiciary Committee Symposium, U.S. House of Representatives

National Association of Secretaries of State

National Conference on Digital Government

Second Conference on Online Deliberation

State Department Speaker/Specialist Tour of Kazakhstan

4. PROJECT HOME PAGE AND TESTBED

For more information:

<http://erulemaking.ucsur.pitt.edu>

The eRulemaking Testbed:

<http://hartford.lti.cs.cmu.edu/eRulemaking/Data.html>

5. ACKNOWLEDGMENTS

We thank personnel at the USDA, DOT, and EPA for providing the public comment data and insights that made this research possible. This material is based on work supported by National Science Foundation (NSF) grants IIS-0429293, 0429102, 0429360, and 0429243. Any opinions, findings, conclusions, or recommendations expressed in this material are the authors' and do not necessarily reflect those of the NSF.

6. REFERENCES

- [1] H. Yang and J. Callan. (2005). Near-duplicate detection for eRulemaking. *Proceedings of the Fifth National Conference on Digital Government Research*. Atlanta, GA.
- [2] H. Yang, J. Callan, and S. Shulman. (Under review). Next steps in near-duplicate detection for eRulemaking. *Proceedings of the Sixth National Conference on Digital Government Research*. San Diego, CA.
- [3] N. Kwon, S. Shulman, and E. Hovy. (Under review). Collective text analysis for eRulemaking. *Proceedings of the Sixth National Conference on Digital Government Research*. San Diego, CA.

SESSION 6C

CRISIS MANAGEMENT 3

Moderator

Jay Kesan, University of Illinois, USA

Titles and Authors

Secure-CITI Critical Information-Technology Infrastructure

Mossé, Daniel; Comfort, Louise; Amer, Ahmed; Brustoloni, José C.; Chrysanthis, Panos K.; Hauskrecht, Milos; Labrinidis, Alexandros; Melhem;Rami; Pruhs, Kirk

E-Government and the Preparation of Citizens for Disasters

Basolo, Victoria; Steinberg, Laura; Gant, Stephen

Secure-CITI Critical Information-Technology Infrastructure

Daniel Mossé, Louise Comfort*, Ahmed Amer, José C. Brustoloni, Panos K. Chrysanthis,

Milos Hauskrecht, Alexandros Labrinidis, Rami Melhem and Kirk Pruhs

University of Pittsburgh

Computer Science Department and *Graduate School of Public and International Affairs

The Secure and robust Critical Information-Technology Infrastructure (S-CITI for short) project aims at providing support to Emergency Managers (EMs) that are faced with management of resources and with decisions before, during, and after emergencies or disasters. Our approach consists of using new and existing sensors to gather data from the field, processing this data to detect and predict emergency/disaster situations, and disseminating this data among the appropriate organizational units. The data flow will be done in a reliable and secure manner and EMs will coordinate actions in a *Virtual Coordination Center (VCC)*, which need not be in a fixed (and thus vulnerable) physical location. The EMs are responsible for indicating what type of data is more valuable, so that S-CITI can display that information appropriately.

There are nine faculty members involved in this effort and five research groups, with several interconnected areas of research. In this paper, we present the "big picture" and then give a brief overview of the current subprojects and results.

The current primary mode of operation in disaster detection and response is through the 9-1-1 system, where humans call in emergencies and the appropriate personnel is dispatched. This detection/response system has proved adequate in many situations, but slow in other scenarios (especially when the humans themselves are involved in the emergency/disaster). Further, counting on reports from the field may be inadequate when communication breakdown exists between the area affected and "the other side" or among personnel responding to the emergency/disaster. A good example of such communication and infrastructure breakdown is the unfortunate levee breaks that happened in New Orleans in 2005 [1].

Our system supports, through information technology, the coordination of usage of existing resources and distribution of the data to different organizational units. The VCC facilitates efficient and quickly coordinated actions to natural and human-caused disasters [2]. The data collected and the actions taken are analyzed through a learning module that will feed post-emergency data into a pre-emergency decision-making module (improving response in the next emergency).

-
- Supported in part through NSF-ITR award ANI-0325353.

The **socio-technical research team (STRT)** combines expertise from data management to social networks, because the issue of Emergency Management is as much a technical challenge as it is a social networks problem. One of the main strengths in our group is the ability to create a connection between the technical and the Emergency Personnel (especially in the Pittsburgh area). Toward the VCC goal, we have developed the *IISIS Executive Dashboard* [3] that provides real-time decision support to practicing EMs during disaster situations.

The STRT developed a preliminary version of a patient-tracking module (see companion powerpoint file), which was demonstrated in a simulated operations exercise planned by the Counter-terrorism Task Force of Southwestern Pennsylvania. This module tracks the information flow from a triage site to a transportation officer to the Receiving Room of an Emergency Medical Department in a hospital. The module supports the exchange of information at all three decision points in a multi-way communication process to enable the timely transport of patients to medical care in the most efficient, informed manner.

Based on this experience and critical evaluation of the STRT, the patient tracking module was revised to include communications among multiple sites of decision making for tracking patients. In essence, the system must enter data and allow monitoring applications to continuously mine data streams for interesting, significant, or anomalous events (e.g., patient data being entered in the system, hospital data being updated, etc). The data must be maintained fresh (not stale data) and the queries must return the most up-to-date values possible. Mechanisms to explore this situation have been reported in [4, 5, 6].

The **ad-hoc network research group (ANReG)** focuses on (a) development of network protocols for energy-efficient data access, (b) accessibility to data even in the presence of disconnected networks, and (c) security of the network.

Because power is a mainstay of ad-hoc mobile networks, which are and will be the type of devices/networks used by emergency personnel (cell phones, PDAs, etc), our group has been investigating networking algorithms (MAC and routing) to minimize the energy consumption in these networks [7]. Our adaptive algorithms have demonstrated significant potential at increasing the longevity of networks used by EMs, relying on optimizing the inquiries to which field officers and volunteers are attempting to respond (e.g., how many beds there are in the ICU of hospital H, what are the traffic conditions to the hospital, etc). Prediction algorithms also contribute to the decrease of the energy and power consumption in these networks, which is a subject of our ongoing investigation. Lastly, our resource management algorithms allow each EM to specify for him/herself the importance of each datum, so that the system can supply the different EMs with the most appropriate data in the most efficient

way. In particular, we will dynamically synthesize databases with new consistency and authentication procedures bound to the type, size, and importance of the data.

The second issue is a new trend in network protocols that abandons the assumption that the network is always connected. We study how couriers (mobile message-forwarding nodes) can enable communication in partitioned networks, carrying messages physically between partitions. Our contribution is a new, cross-layer routing approach based on the observation that orders from an EM (leader sending tasks to responders) also controls responders' mobility and thus their ability to forward messages. We schedule responders' movement considering both their tasks and network needs. Our simulations demonstrate performance benefits of our approach in a variety of scenarios [8]. We have also developed a prototype EOC-like testbed with sensors and mobile nodes for experimentation in this area.

The third issue is that, in mobile devices, not only energy and connectivity are important factors, but the issue of security is especially important because the network infrastructure may be under attack. For that, we are investigating two problems. First, when couriers that receive messages from or send messages to the EOC are being jammed. We have modeled this scenario through a Markov Model, in which the state of the couriers can be blocked, free to receive, or moving (to another location to avoid jamming). Our simulation and modeling results allow for determining the exact break-down point, that is, when the EOC should either add couriers, or suffer from lack of connectivity [9]. Our second aim is to provide security at the level of the network, not allowing messages to be forged, intercepted, deleted, or inserted by attackers. Clearly, our protocols respect the different organizational units and the privacy of data.

The *intelligent monitoring and diagnostics group (IMDiG)* has focused on modeling static and dynamic dependencies in large distributed systems with continuous and discrete random variables, and efficient optimization of control activities in such distributed systems. Because transportation systems are one of the key elements of the emergency response system, one of the goals of our work is to build statistical models that let us represent spatial correlation patterns among traffic variables (speed, volume, throughput, etc) over the network under different conditions (free-flowing normal, traffic congestion). Such models support EMs under a variety of distress conditions. For example, understanding congestion patterns and traffic predictions would allow the EMs to better route/dispatch emergency resources (ambulance, fire, repair units) in a way that reduces response time.

We have obtained two years worth of traffic data from over 100 sensors placed on major Pittsburgh highways and have investigated statistical properties of the data and relations among sensors themselves. We have built initial multivariate models that can lead to, for example, prediction of different events. A proof-of-concept algorithm was developed that successfully detects anomalous days from vehicular count data. Our concrete goal is to develop an online system that predicts with high probability a possible accident before it happens, by simply using inferences on the data obtained from the sensor network in place.

Finally, the *advanced data management technologies laboratory (ADMT)* in collaboration with the *storage research group (SRG)* have produced a wide range of data management

algorithms, focusing in quality of data (QoD) and quality of service (QoS), in the presence of resource constraints, which characterize emergency response environments. For example, at the point of generation of data (e.g., from networks of sensors or mobile devices), we have proposed energy-efficient data acquisition techniques [10], which combine in-network processing [11] and in-network, data-centric storage [12]. To effectively propagate data within such sensor networks, we proposed semantic-based, multi-criteria, self-optimizing algorithms for constructing efficient and robust routing topologies [13]. Data in the emergency management domain typically come in the form of data streams that need to be continuously monitored for interesting or anomalous events. Towards this, we have proposed QoS- and QoD-aware techniques for processing of continuous queries [4, 5] and implemented some of these ideas in our prototype Patient Tracking module, mentioned above. The produced data is disseminated to mobile users, such as first responders, who utilize our proposed energy-efficient mobile caching and energy-safe prefetching algorithms, which are self-optimizing for changing workloads [14].

REFERENCES

- [1] Comfort, L.K. *Communication, Coherence, and Collective Action*. Public Works Management & Policy., Sep. 2006.
- [2] Berfield, A., Chrysanthis, P.K., Labrinidis, A.. *Automated Service Integration for Crisis Management*, 1st Int'l ACM Workshop on Databases in Virtual Organizations, 2004.
- [3] IISIS prototype, <http://www.iisis.pitt.edu>
- [4] Sharaf, M., Chrysanthis, P.K., and Labrinidis, A. *Preemptive Rate-based Operator Scheduling in a Data Stream Management System*. IEEE AICCSA, 2005.
- [5] Sharaf, M., Labrinidis A., Chrysanthis P.K., and Pruhs, K. *Freshness-Aware Scheduling of Continuous Queries in the Dynamic Web*. ACM WebDB, 2005.
- [6] Qu, H., Labrinidis, A., and Mossé, D. *UNIT: User-centric Transaction Management in Web-Database Systems*. IEEE ICDE 2006.
- [7] Gobriel S., Melhem, R., and Mossé, D. *BLAM: An Energy-Aware MAC Layer Enhancement for Wireless Adhoc Networks*. IEEE WCNC, 2005.
- [8] Brustoloni, J., Khattab, S., Santamaria, C., Smyth, B. and Mossé, D. *Integration of Application-Layer Scheduling and Routing in Delay-Tolerant MANETs*. Pitt CSD T.R., 2006
- [9] Khattab, S., Mossé, D., Melhem, R. *Honeybees: Combining Replication and Evasion for Mitigating Basestation Jamming in Sensor Networks*. WPDRTS, 2006.
- [10] Sharaf M., Beaver J., Labrinidis A., Chrysanthis, P.K. *Balancing Energy Efficiency and Quality of Aggregate Data in Sensor Networks*, The VLDB Journal, Dec. 2004.
- [11] Xia, P., Chrysanthis, P., Labrinidis, A.. *Similarity-Aware Query Processing in Sensor Networks*. WPDRTS, 2006.
- [12] Aly, M., Morsillo, N., Chrysanthis, P.K., and Pruhs, K.. *Zone Sharing: A HotSpots Decomposition Scheme for DataCentric Storage in Sensor Networks*, DMSN, 2005.
- [13] Li Q., Beaver J., Amer A., Chrysanthis P., Labrinidis A., Santhanakrishnan G. *Multi-Criteria Routing in Wireless Sensor-Based Pervasive Environments*, Journal of Pervasive Computing and Communications, Dec. 2005.
- [14] Larkby-Lahet J., Santhanakrishnan G., Amer A., Chrysanthis, P.K. *STEP: Self-Tuning Energy-safe Predictors*, ACM/IEEE MDM Conference, 2005.

E-GOVERNMENT AND THE PREPARATION OF CITIZENS FOR DISASTERS

Victoria Basolo
University of California, Irvine
202 Social Ecology I
Irvine, CA 92697-7075
1-949-824-3521
basolo@uci.edu

Laura Steinberg
Tulane University
Civil and Environmental Engineering
New Orleans, LA 70118
1-504-862-3254
lauras@tulane.edu

Stephen Gant
Computer Sciences Corporation
440 Fair Oaks Circle
Chapel Hill, NC 27516
1-919-942-3337
stephen.gant@ieee.org

Keywords

Risk Communication, Hazard Preparedness, Web, Cities, Free-listing

1. INTRODUCTION

Local community web sites are potentially powerful tools to increase citizens' knowledge about environmental risks and hazard preparedness [1]. However, there is virtually no research on the development of local governments' web sites for hazard preparedness or the usability of this information technology by community residents.

This research investigates the use of the World Wide Web (web) by local governments and community residents for delivering and receiving risk and preparedness information. The study areas are incorporated cities in Los Angeles (n=88) and Miami-Dade Counties (n=31); these regions were chosen because they are subject to major natural disasters, as well as human-induced disasters. The research uses mixed methods including quantitative analysis of the city web sites in both study regions, a sample telephone survey of residents in each region, a user experiment using pre-test, post-test surveys and user logs, and qualitative case studies of local government decision making concerning the city websites, especially decisions on the degree and content of risk and hazard preparedness information on the municipal site.

2. CURRENT PROJECT ACTIVITIES

The researchers constructed a core set of search terms, known as a synonym ring, for disaster topics based on starting terms/phrases from the FEMA web site, the WordTrack.com meta search log service, and free-listing data gathered from residents in Southern California and the New Orleans areas (see Figure 1 for key term/phrase identification process).

The researchers constructed a snapshot of the 119 websites by spidering them with HTTrack and saving the cities' pages to a local file system, first in June 2005 (Pre-Katrina/Rita) and then in January 2006. The researchers applied automated tools to measure 14 facets of site findability used by the major search engines to determine page rank [2].

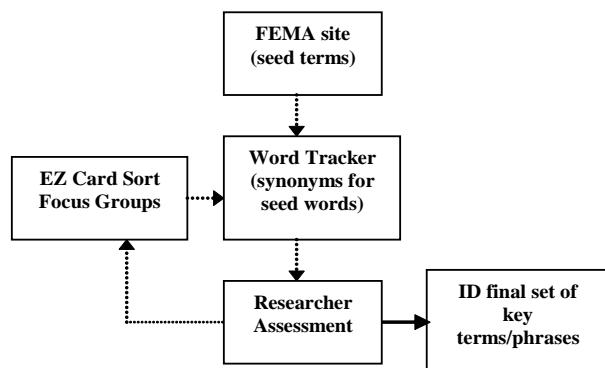


Figure 1. Key term identification process

Table 1 shows results for three of the indicators used in spidering the municipal web sites. Less than half of the city sites in the study areas had the disaster keywords in their homepage, a link from the homepage to content with keywords, or keywords in HTML link tags anywhere on the site. These results indicate many cities in the study areas are not using their website to communicate hazard risk and preparedness information.

Table 1. Cities with one or more keywords/links, 2005 & 2006

Variable	Los Angeles County (n=88)				Miami-Dade County (n=31)			
	2005		2006		2005		2006	
	#	%	#	%	#	%	#	%
Keywords (homepage)	30	34.1%	27	30.7%	8	25.8%	9	29.0%
Link homepage to content	33	37.5%	34	38.6%	13	41.9%	14	45.2%
Keywords (HTML link tags)	37	42.0%	35	39.8%	13	41.9%	15	48.4%

Human coders evaluated 20 measures of usability for the home pages of the study cities in 2005 and are currently coding the 2006 versions of the city web sites, including the existence of disaster preparedness/information on the home page and direct links to underlying content [3]. City home pages will be compared for change between 2005 and 2006. Finally, the coders will evaluate each website's disaster preparedness content on 57

measures developed by the research team. The individual measures will be used to create a disaster preparedness communication index. The researchers have disaster preparedness content information from 90% of the study cities for 2005; however, 42% of the total sites have no content (see Table 2). The content coding for 2006 (and reliability check of 2005 web sites) will be completed by late March 2006. Content change from 2005 to 2006 will be assessed with the expectation that disaster preparedness content will have increased between the two periods. An overall site rating for the 2006 versions of the web sites will be determined using the home page and content evaluations and these sites will be ranked from excellent to poor.

Table 2. Website content, 2005

Descriptions	Sites	
	#	%
Contains disaster preparedness information for citizens (may also contain links to external sites)	38	32%
Minimal text, reference to external sites	12	10%
Disaster response training/capability description (no links or other preparedness artifacts)	7	6%
No disaster preparedness information on site	45	38%
No city website	5	4%
Unable to save (problems with Flash-based pages or portal-driven sites)	12	10%
Total:	119	100%

3. RESEARCH CONTRIBUTIONS

The website spidering and content coding provide a good understanding of the use of municipal websites to communicate hazard risk and preparedness. As discussed above, many cities do not appear to be using their municipal website for these purposes. This contribution alone is significant because it reveals an opportunity for enhancing hazard risk and preparedness communication for residents of communities.

The early research results also suggest that the disasters of the 2005 hurricane season had little effect on the content of municipal websites. Difference of means tests on the three indicators discussed in the previous section indicate that only one, keywords in HTML link tags anywhere on the site, increased significantly and only for the cities in Miami-Dade County. However, the distributions of the indicators are non-normal and therefore, conclusions based on the t-tests must be made cautiously.

Finally, a preliminary logistic regression analysis, based on the 2006 spider data, indicates that city population size has a statistically significant, but very small effect, on whether cities have at least one target keyword in HTML link tags. In other words, these results suggest that city size is not an important predictor of this indicator. Moreover, this analysis shows cities in Miami-Dade County are approximately three times more likely than cities in Los Angeles County to have at least one target keyword in HTML link tags versus no target keywords in HTML link tags, controlling for city population size.

4. RESEARCH PLAN AND CHALLENGES

The progress of the research was slowed due to the severity of the 2005 hurricane season. The sample survey of residents was delayed, but is now on a fast track schedule. This survey is designed to assess residents' perception of risk for various hazards, determine the emergency preparedness of residents within the study regions, and their use of local government websites, including for hazard information. Currently, the questionnaires are in draft form to be pretested and are expected to be in the field by March 15, 2006. The Miami questionnaire will emphasize hurricane risk and the Los Angeles survey will emphasize earthquake risk. The survey will take approximately one month to complete. The user experiments use information from the coding phase of the research. Therefore they have been scheduled for May 2006. The content coding will identify the best and worst preparedness sites among our sample of cities based on our established criteria (i.e., index score). The experiment involves 60 subjects, randomly selected into one of three groups. Each group will be given different instructions including: a short description of the "problem" (you want information on disaster preparedness), specific sites to visit (best for one group; worst for second group; and free searching for third group). Pre- and post-tests of attitudes toward risk and emergency preparedness will be given to all the subjects to determine if reading the web sites affects their perceived likelihood of preparing for a disaster event. These short tests also will determine if groups varied in their assessment, as users, of the sites based on the quality ranking established by the coding analysis. In addition, subjects will keep user logs of their perceptions of the searching process and the visited web sites for qualitative text analysis by the research team. Finally, in late May through June 2006, members of the research team will visit selected cities (with websites ranked high, as well as low, for their risk and preparedness information) in the Miami and Los Angeles regions. Interviews will be conducted with city officials to reveal the decision making processes for city website content, especially for risk communication and hazard preparedness.

The project faced significant challenges in 2005, most notably the interruption of the project due to the devastating hurricane season. The dislocation of the Co-PI from her home institution in New Orleans, as well as delays on the project related to work-in-progress at the time of Hurricane Katrina, put the project off schedule. However, several tasks have been implemented and others rescheduled to place the project back on track for a completion date of September 30, 2006.

5. REFERENCES

- [1] California LAO. 2001. E-Government in California: Providing Services to Citizens Through the Internet. Available at: http://www.lao.ca.gov/2001/012401_egovernment.html.
- [2] Sullivan, D. (Editor). Search Engine Placement Tips (Web page), 2002. Available at <http://www.searchenginewatch.com/resources/article.php/2168021>.
- [3] Koyanl, S., Bailey, R., and Nall, J. (Editors). Research-Based Web Design & Usability Guidelines [Web Page], National Cancer Institute, 2003. Available at <http://usability.gov/pdfs/guidelines.html>.

SESSION 7A

PARTICIPATORY DESIGN AND MEDIATION

Moderator

Marianne Lykke Nielsen, Royal School of Library and Information Science, Denmark

Titles and Authors

*Developing a Youth-Services Information System for City and County Government:
Experiments in User-Designer Collaboration*
Zappen, James P.; Adali, Sibel; Harrison, Teresa M.

Policy Through Software Defaults
Shah, Rajiv C.; Kesan, Jay P.

*Experimental Application of Process Technology to the Creation and Adoption of Online
Dispute Resolution*
Sondheimer, Norman K.; Katsh, Ethan; Rainey, Daniel; Osterweil, Leon J.

Developing a Youth-Services Information System for City and County Government: Experiments in User-Designer Collaboration

James P. Zappen

Rensselaer Polytechnic Institute
Department of Language, Literature,
and Communication
Troy, New York
518-276-8117
zappenj@rpi.edu

Sibel Adali

Rensselaer Polytechnic Institute
Department of Computer Science
Troy, New York
518-276-8047
sibel@cs.rpi.edu

Teresa M. Harrison

University at Albany, SUNY
Department of Communication
Albany, New York
518-442-4883

harrison@albany.edu

ABSTRACT

Research on user participation in computer applications, including digital-government applications, emphasizes the need to engage users as collaborators or partners in software-design processes. The Connected Kids youth-services information system for city and county government is a product of ongoing experiments in user-designer-programmer collaboration in the development of system specifications and prototypes and a functional working model, still in process of modification in response to user needs and creative user-designer-programmer innovations.

Categories and Subject Descriptors

J.4. [Computer Applications]: Social and Behavioral Sciences –
Communication

General Terms

Design, Theory, Experimentation

Keywords

Digital Government, City Government, County Government, Information System, Youth Services, Cooperative Design, Participatory Design, User-Centered Design, Co-Design, Communication, Collaboration

1. INTRODUCTION

The development of a youth-services information system for city and county government illustrates the potential for creative collaborations by users, designers, and programmers through exchanges of information and perspectives about information-technology resources and potentials, on the one hand, and user needs and creative on-site innovations, on the other. Research on user-centered design and user-developer co-design emphasizes the need for user engagement in design processes extending from the development of system specifications, prototypes, and working models to the incorporation of user innovations during and after implementation of information systems and other information-technology applications.

This research originated in Scandinavia [2, 3, 6] and has been adopted in the United States as a set of principles and practices various described as contextual [1], cooperative [2, 3, 6], participatory [9], and user-centered design [8] and as co-design

between users and developers [10]. More recently, research on the development of information-technology applications for digital government has acknowledged the important role of users, who *enact* information technologies as they interpret, implement, and use these technologies in the context of their own organizations [5]. These movements foregrounding the role of users in the design and implementation of information technologies are supported by technical developments that permit and encourage user modifications of system components and functions.

The Connected Kids Project (<http://www.connectedkids.info/>, November 25, 2005) is a youth-services information system for Troy and Rensselaer County, New York, currently active but still in the process of development [7, 11]. As a digital-government application for local governments and local youth-services organizations, Connected Kids has a unique opportunity to observe experiments in the various interactions and modes of cooperation between and among participants throughout the development and implementation of an information system. At the level of local government, organizational users, designers, and programmers interact regularly, with the consequence that new developments in software technology, such as APIs (Application Program Interfaces), mix and merge with end-user innovations in processes that are mutually reinforcing and enriching.

In this report, we review some of the research on user-centered design and user-developer co-design and present an overview of the Connected Kids Project, an account of its development, and a description of some recent experiments in what we call *user-designer-programmer collaboration*, to emphasize the balance of interests and expertise across the spectrum of development activities that engage organizational users, interface designers, and computer programmers.

2. FROM USER-CENTERED DESIGN TO USER-DESIGNER COLLABORATION

Research on the design and use of computer applications shows a shift from user-centered design to user-designer collaboration, extending throughout the development and implementation cycles and including innovations introduced in the context of use [1-3, 6, 8-10]. These shifting perspectives blur the distinction between users, designers, and programmers, as system users become designer-programmers and designer-programmers become system users.

2.1 User-Centered Design

Researchers have brought a variety of user-centered or user-oriented approaches and methods to the design of computer applications, encompassing the full range of design activities, from the development of concepts and specifications to the development of models and prototypes through testing, implementation, and use and application of end products [1-3, 6, 8-10]. Early work originating in Scandinavia promoted an ideal of “cooperative development—full participation by both developers and users” on the basis of cost-benefit advantages, enhanced democracy in workplace practices, and, not least, improved product quality [2, pp. 157-58; 3, pp. 130-31; 6, pp. 79-80].

Extending and generalizing this early work, Beyer and Holtzblatt propose methods of contextual inquiry and design by which developers engage end users in collaborative partnerships for the purpose of exploring and modeling workplace practices, developing representations of customer populations, creating innovative designs, and refining designs through a process of iterative prototyping [1]. In contextual inquiry and design, users are “partners,” “collaborators,” and “co-designers” in the design process [1, pp. 37, 51-56, 371-77, 397-98]. Johnson presents a similarly “user-centered” view of the design of computer documentation [8, pp. 12-16]. As distinguished from traditional system-centered and user-friendly views, Johnson’s user-centered view of computer documentation takes into account the localized context of users’ activities, their choice of media and activities (doing, learning, or producing), and the social interactions and negotiations by which these contexts, media, and activities are incorporated into documentation products [8, pp. 122-36].

2.2 User-Developer Co-Design

More recent research extends and transforms the concept of user-centered design into a concept of user-developer co-design that recognizes the reciprocities between users and designer-programmers in design activities [10]. Spinuzzi maintains that traditional approaches to user-centered design are based upon a user-as-victim, designer-as-hero trope, according to which users are victims waiting to be rescued by skilled designers [10, p. 4]. As an alternative to these traditional approaches, Spinuzzi offers a concept of co-design that recognizes the contributions of both devious and wily users and skillful designers: “Trained designers can avoid common pitfalls of workers’ homegrown solutions, which tend to be of the chewing-gum-and-bailing-wire variety. Workers produce solutions that are devious, wily, and cunning, but often these solutions do not involve a deep understanding of the system Workers produce solutions that work—but often they do not produce solutions that work well *by their own criteria*, and often those solutions are not promulgated so that other workers can take advantage of them” [10, pp. 19-20]. Given the potentially creative and innovative contributions of users, Spinuzzi advocates a “decentralized” approach to design that permits and encourages user modifications: “Trained information designers can contribute much to the emergent innovations of workers, not by replacing those innovations with centralized solutions, but by helping to design systems that workers can modify” [10, pp. 4-5, 222-23].

Spinuzzi’s view of user-developer co-design is partially realized in new and emerging software technologies that encourage adaptations and modifications by both skilled developers and end users. One example of such a technology is the Google Web APIs

service, which permits skilled software developers to “query billions of web pages directly from their own computer programs” in their favorite development environments, such as Java, Perl, or Visual Studio [<http://www.google.com/apis/>, November 25, 2005]. Another example is the EUSES (End Users Shaping Effective Software) Consortium (<http://eusesconsortium.org/>, November 25, 2005), a collective of universities organized to effect a fundamental paradigm shift in software technology by exploring the feasibility of bringing the benefits of rigorous software engineering methodologies to end users (About EUSES, November 25, 2005). These emerging software technologies parallel the more modest innovations of ordinary users but promise, in the long term, to support and encourage these innovations as the technologies become more flexible and adaptable and as increasing numbers of users become more technically skilled programmers. In relatively small-scale projects, in which users and developers regularly interact, these emerging software technologies and user innovations mutually enrich and support each other.

2.3 Digital-Government Applications and User-Designer-Programmer Collaboration

Research in digital-government applications reflects these trends and tendencies in user-centered design and user-developer co-design. Traditional digital-government research has emphasized the need to deploy information technology to deliver information more efficiently: “Stimulated in large part by widespread adoption of the Internet and the associated phenomenon of electronic commerce, a broad consensus has emerged in the past several years that government at all levels can exploit IT to deliver information and services more efficiently and to make improvements in other functional areas” [4, pp. 29-30]. These functional areas include satisfying customer-service expectations, increasing the efficiency and effectiveness of government operations, providing effective online access to information and transactions, and increasing participation in government, among others [4, p. 30].

Fountain, however, recognizes that information technology does not stand in a simple linear relationship to its users [5]. Rather, information technology is enacted by its users in the context of use: “Individuals and organizations enact information technology by their interpretation, design, implementation, and use of it in their organizations and networks.” [5, p. 89]. These “enacted information technologies” differ from and may represent only a limited subset of the capabilities offered by objective technologies—the Internet, hardware, software, etc. [5, p. 98]. Nonetheless, the outcomes of enacted technologies are “multiple, unpredictable, and indeterminate” [5, p. 98]. We imagine, therefore, that these enacted technologies have the potential to supplement, modify, or refine the objective characteristics of existing information systems. At the local level, this process of user enactment can be especially rich and complicated since local governments are more likely to have direct and frequent interactions with users than their counterparts at the federal and state levels. In the rest of this report, we illustrate this process of user enactment, taking the term *user* in the broad and complex sense that encompasses users, designers, and programmers in their shifting and varying contexts and functional roles and characterizing their relationship as a *user-designer-programmer collaboration*.

3. THE CONNECTED KIDS PROJECT

The Connected Kids Project is currently serving city and county governments, youth-services organizations, public and private schools, and families and children in Troy and Rensselaer County, New York. The Connected Kids information system includes a database of youth-services information with sophisticated search capabilities, a distributed data-input function, and separate interfaces for parents, teens, and kids—all accessible via the World Wide Web (Figure 1, Connected Kids Home Page); galleries of children’s artwork and photos (Figure 2, Connected Kids Galleries)—also accessible via the WWW; and a distribution system that extends the information system and galleries to low-income families at the Troy Housing Authority.



Figure 1. Connected Kids Home Page

The Connected Kids Home Page shows the points of entry for parents, teens, kids, and youth-services organizations and for the galleries and general-information pages. The parents, teens, and kids pages permit searching by key word or browsing by activity category, sponsoring organization, age, gender, time frame, or calendar date. The organization pages provide for registration, creation of an organization profile, data entry for organizational activities, image editing and storage, and the creation of a set of customized web pages. The organization pages offer an easy-to-use interface, with simple fill-in-the-blanks and copy-and-paste operations.



Figure 2. Connected Kids Galleries

The Connected Kids Galleries include galleries of children’s artwork and photos and a separate section with information about

local art and history. The children’s artwork and photos are presented in a variety of formats, including user-operated slide shows with images from a number of youth-services organizations, with the requisite permissions, including signed Connected Kids permission forms for all photographs of children.

The Connected Kids distribution system includes computer and networking installations at six sites at the Troy Housing Authority, four with Linux setups and two with Windows setups. Four of the sites, including one for senior citizens, currently have Internet connections. In addition, the distribution system includes instructional support and periodic help and troubleshooting.

4. DEVELOPMENT OF THE CONNECTED KIDS INFORMATION SYSTEM

The early development of the Connected Kids information system engaged more or less standard approaches and methods described in the research on cooperative and participatory design: focus groups to develop system specifications, participatory-design and user-testing sessions to assess system prototypes, training sessions, and standard evaluation procedures [7, 11]. In our initial focus groups to develop system specifications, we presented mock-ups showing possible functionalities and uses of the system [7, 11, pp. 361-62]. In response, organizational users expressed concerns about what seemed to them to be an overemphasis upon listings of calendar events as opposed to descriptions of organizational programs and services, problems of duplicate data entry for both large and small organizations, the lack of a web presence for many small organizations, and the need to engage parents and children in the development of system specifications [7, 11, p. 362]. We developed a paper-based model of the system and a working system prototype on the basis of these responses and then conducted participatory-design and user-testing sessions to test and further refine the system specifications [7, 11, pp. 362-63]. During the participatory-design sessions, organizational users reconfirmed their original responses, described how the system would function in their organizations, and identified a variety of user needs, including the need to serve a growing Spanish-speaking population, to provide legal services, to support neighborhood activism, to ensure system security, and to engage parents and children in the design process [7, 11, pp. 362-63]. We have been able to address some but not all of these needs.

As an immediate priority, we conducted another series of focus groups with parents and children to further refine system specifications and to develop specifications for the WWW interface [11, pp. 364-66]. Because we suspected that relatively inexperienced computer users might have difficulty grasping the functions of a WWW-based information system, we presented working illustrations of search and browse operations for parents and dynamic WWW interfaces for children, selected by the children themselves. We did not ask parents and children to describe or explain possible system functions but instead asked parents to describe how they find activities for their children and for themselves and asked children to describe the kinds of web resources they use and enjoy [11, p. 359]. In response, both parents and children offered descriptive scenarios rather than explanations of how an information system might function [11, pp. 364-66]. Parents, for example, described challenges ranging from finding suitable activities for their children to keeping them productively occupied to sharing information about the quality of

available activities [11, pp. 364-66]. Children, not surprisingly, emphasized the need to develop a dynamic and visually attractive and engaging interface. On the basis of these responses, we developed the WWW interface and galleries shown in Figures 1 and 2. We are currently evaluating both the system and the interface, including the galleries, and developing further enhancements, including a new map interface, on the basis of these evaluations.

5. EXPERIMENTS IN USER-DESIGNER-PROGRAMMER COLLABORATION

These more recent developments reflect the underlying principle of co-design or what we call *user-designer-programmer collaboration* since they are based upon ongoing interactions between users—primarily organizational users—designers, and programmers. These developments include new data-entry and display functions, new search functions and the map interface, and new gallery and help functions.

5.1 Data-Entry and Display Functions

The development of the data-entry and display functions is based in part upon the innovations of organizational users and the efforts of designer-programmers to build these innovations into the system. These innovations are the “homegrown” or “chewing-gum-and-bailing-wire” solutions to organizational problems and needs that designers need to stabilize and thus make accessible to other users [10, p. 19].

Most of the Connected Kids organizational users are comfortable with computer technology, and some have limited programming or basic coding experience (e.g., html or xhtml). One of these organizational users—unexpectedly but almost immediately upon implementation of the system—bypassed the system’s image editing and storage function and wrote a small piece of code that pointed to a logo on the organization’s own server and thus inserted it into the descriptive field for one of the organization’s activities (Figure 3, Activities Display Page with Organizational Logo). Other users expressed interest in similarly inserting logos for their organizations, and one user inserted a very large graphical image, straining system capacity and display functions.



Figure 3. Activities Display Page with Organizational Logo

In response to these developments, we placed limits on image size and implemented a resize feature in the image-editing function. In addition, we recognized the utility of inserting html code into the activities descriptive fields and took advantage of this utility on

other occasions. We recalled that organizational users had expressed concern about the potential problem of duplicate data entry, and we noted the difficulty of entering, or re-entering, large and complex documents, such as school or sports calendars, into the system. In one instance, therefore, we took advantage of the html utility to point to these documents, housed on a local server (Figure 4, Activities Data-Entry Field with HTML Code).

Figure 4. Activities Data-Entry Field with HTML Code

This “chewing-gum-and-bailing-wire” solution had very limited value, however, since it depended upon users’ ability to write html code (and only a few could write even simple code) and upon the availability of documents on local servers. We therefore built a document-upload function into our customized web pages, with the capability of converting standard MS Word documents into .pdf files (Figure 5, Customized Web Page). These pages, currently operational, include separate pages (under More About Us) for the organization profile, additional information (the uploaded documents), and staff and location listings.



Figure 5. Customized Web Page

The uploaded documents are accessible through each organization’s customized web pages and also its activity listings. For the Troy Housing Authority, the uploaded documents include housing application and survey forms, newsletters, and activity schedules, accessible electronically at the four sites with Internet connections and, eventually, at others, as the connections become available.

5.2 Search Functionality and Interface

The new search functions and map interface are designer-programmer innovations that draw upon recent developments in

software technology aimed at the creation of “decentralized” systems that user-designer-programmers can adapt to their own purposes and needs—such as the Google Web API technology and the EUSES Consortium projects described above [10, pp. 222-23]. We are using the Google Web API technology to develop our own application in the form of enhanced search capabilities via a map interface (Figure 6, Mockup of a Map Search Interface).

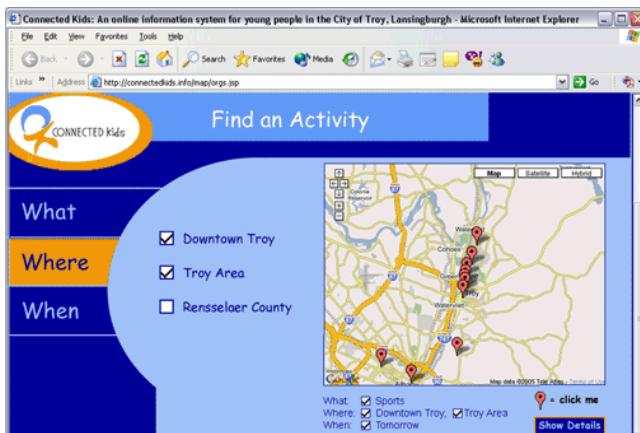


Figure 6. Mockup of a Map Search Interface

The Google Web API permits integration of the Google search capability with local applications, such as the Connected Kids information system. Recognizing the preferences of both parents and—especially—children for a dynamic and interactive search interface, we are working to enhance both the search capability and ease of use of the search interface. For our application, currently in beta, we are using the Google API as a search mechanism that permits users to search by triangulating kind of activity (what), location (where), and date and time (when) or by selecting an organization by name or by map location. To ensure ease of operation, we have created an interface to permit users to select one or more of the what, where, and when choices via simple point-and-click operations and display the specified results. We have also adapted the Google map interface to permit users to select an organization by name or map location and search for activities via the map. Using this map interface, the user can zoom in on a selected location, scroll the map in any direction to survey neighboring locations, or select About Us or Activities to go directly to the organization profile or activities pages to view the results.

To further extend and enhance the search capability, we are currently exploring the potential of the Google Web API to search the webs of organizations linked to the organization profile and activities pages and display the results along with the results for activities currently entered into the system. This functionality would extend the reach of the system, of course, but would also complement existing system capabilities since organizations could continue to enter documents through the customized web.pdf function but would not have to enter documents currently displayed via their own webs.

5.3 Gallery and Help Functions

In addition to the system search functions, other system enhancements are also possible via user-designer-programmer collaborations. For the Connected Kids Galleries, for example, we have engaged in a series of exchanges with organizational users to

enhance the presentation of children’s artwork and photos. After a series of experiments, we settled upon a slideshow as the best method of displaying the images since it permitted the maximum of user control and manipulation of the images. To further enhance the image displays, we created a selection of colorful and attractive backgrounds for the slideshows. In the first iteration, we affixed the images to the background, with a consequent increase in load time and limited flexibility in the implementation of the displays. In the next and most recent iteration, we created separate layers for the image, background, and navigation components of the display, so that organizational users could create their own backgrounds for their images (Figure 6, Gallery Display with Layers).

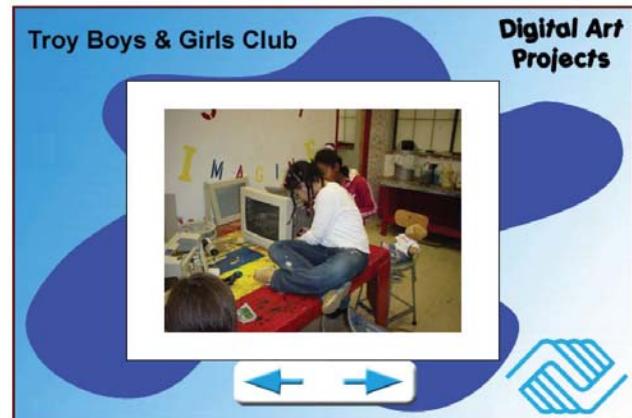


Figure 6. Gallery Display Page with Layers

For maximum control and flexibility for organizational users, we could, in principle, create an interactive interface that would permit users to enter their backgrounds and images for themselves. But such a development would require a considerable investment in technology, including the increased security required for the gallery displays.

For the Connected Kids distribution system, in addition to the computer reconstruction, networking, and maintenance, we have offered instructional and help and troubleshooting support. In one instance, as an experiment, we created a wiki as a collaborative mechanism to ensure maximum technical support on short notice (Figure 7, E-GroupWare Wiki Gmail Tutorial).

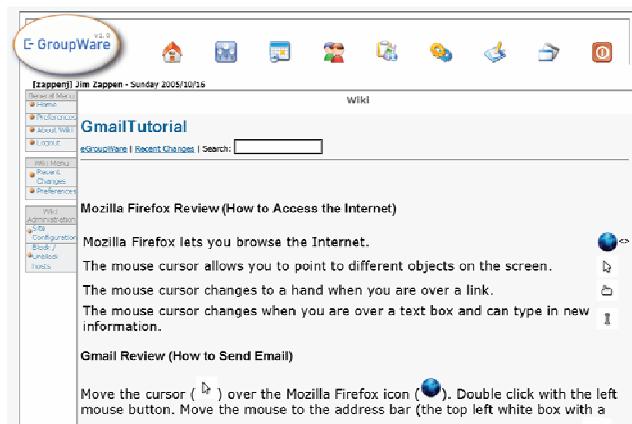


Figure 7. E-GroupWare Wiki Gmail Tutorial

In practice, the wiki permits us to respond to telephone or email requests for technical support and assistance by updating tutorials and other system helps remotely and posting updates within hours or even minutes upon receipt of a request. However, we have not yet taken the adventuresome and perhaps risky step of permitting either senior citizens or children to contribute directly to the wiki, even though the technical capability is readily available.

6. CONCLUSION

Recent research on user-centered design and user-developer co-design, principles of user enactment of digital-government applications, and emerging software technologies such as APIs recognize and support user innovations and enactments of information-technology applications. The development of the Connected Kids information system illustrates the potential for user-designer-programmer collaborations in the context of local government, where the interactions between system users and developers are frequent and often creative. For our application, the results of these interactions impact virtually all aspects of the system, from the search functions to the search and gallery displays to even the most basic image display and help functions.

7. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 0091505. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Connected Kids is supported by the National Science Foundation, the 3Com Urban Challenge Program, the Rubin Community Fellows Program, Rensselaer's School of Humanities and Social Sciences, and other organizations. We are grateful to Rensselaer graduate students Maureen Duffy, Sin-Hwa Kang, Stacy Newman, and Matthew Novak for assistance with the development of the Connected Kids interface and galleries and for discussions about the ideas in this report. We are grateful to the Lansingburgh Central School District, the Troy Boys & Girls Club, the Troy Housing Authority, and other organizational users for their creative innovations to the system.

8. REFERENCES

- [1] Beyer, H., and Holtzblatt, K. *Contextual Design: Defining Customer-Centered Systems*. Series in Interactive Technologies. Morgan Kaufmann Publishers, Academic Press, San Diego, CA, 1998.
- [2] Bødker, S., Grønbæk, K., and Kyng, M. Cooperative design: techniques and experiences from the Scandinavian scene. In *Participatory Design: Principles and Practices*. D. Schuler and A. Namioka, eds. Lawrence Erlbaum Associates, Hillsdale, NJ, 1993, 157-75.
- [3] Bødker, S., and Grønbæk, K. Users and designers in mutual activity: an analysis of cooperative activities in systems design. In *Cognition and Communication at Work*, Y. Engeström and D. Middleton, eds. Cambridge University Press, Cambridge, UK, 1998, 130-58.
- [4] Committee on Computing and Communications Research to Enable Better Use of Information Technology in Government, Computer Science and Telecommunications Board, Division on Engineering and Physical Sciences, National Research Council. *Information Technology Research, Innovation, and E-Government*. National Research Council, Washington DC, 2002.
- [5] Fountain, J. E. *Building the Virtual State: Information Technology and Institutional Change*. Brookings Institution Press, Washington DC, 2001.
- [6] Grønbæk, K., Grudin, J., Bødker, S., and Bannon, L. Achieving cooperative system design: from a product to a process focus. In *Participatory Design: Principles and Practices*. D. Schuler and A. Namioka, eds. Lawrence Erlbaum Associates, Hillsdale, NJ, 1993, 79-97.
- [7] Harrison, T. M., and Zappen, J. P. Methodological and theoretical frameworks for the design of community information systems. *Journal of Computer-Mediated Communication*, 8, 3 (Apr. 2003).
- [8] Johnson R. R. *User-Centered Technology: A Rhetorical Theory for Computers and Other Mundane Artifacts*. Studies in Scientific and Technical Communication. State University of New York Press, Albany, NY, 1998.
- [9] Schuler, D. and Namioka, A., eds. *Participatory Design: Principles and Practices*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1993.
- [10] Spinuzzi, C. *Tracing Genres through Organizations: A Sociocultural Approach to Information Design*. MIT Press, Cambridge, MA, 2003.
- [11] Zappen, J. P., and Harrison, T. M. Intention and motive in information-system design: toward a theory and method for assessing users' needs. In *Digital Cities 3: Information Technologies for Social Capital*, P. van den Besselaar and S. Koizumi, eds. Lecture Notes in Computer Science, vol. 3081. Springer-Verlag, Berlin, Germany, 2005, 354-68.

Policy Through Software Defaults

Rajiv C. Shah

University of Illinois
College of Law
504 E. Pennsylvania Ave

rshah@a5.com

Jay P. Kesan

University of Illinois
College of Law
504 E. Pennsylvania Ave.
(217) 333-7887

kesan@uiuc.edu

ABSTRACT

As part of digital government, policymakers are increasingly considering the use of software to influence societal concerns such as privacy, freedom of speech, and intellectual property protection. A necessary step is deciding what the settings should be for software. In this paper, we build upon work in computer science and behavioral economics to argue how defaults in software should be set.

Categories and Subject Descriptors

K.4.1 [Computers and Society]: Public Policy Issues-Regulation, Privacy

General Terms

Legal Aspects

Keywords

Regulation, law, code, defaults, digital government

1. INTRODUCTION

Research in digital government typically focuses on how information technologies affect society. This paper approaches the digital government issues from the opposite perspective. We argue that policymakers can harness information technologies to increase society welfare. In this paper, we describe how policymakers can influence behavior by manipulating default settings in software. Changes in default settings can materially improve fundamental societal concerns such as privacy and security.

There is a well developed recognition that information technologies can influence or regulate behavior affecting fundamental societal issues, such as privacy [16]. Our previous work has developed this idea by putting forth a theoretical

framework for how information technologies affect society and how, in turn, society influences the design of information technologies [14]. A crucial part of this framework was the recognition that there are certain characteristics of information technologies that influence behavior [25]. These governance characteristics are analogous to “knobs and levers” that policymakers can manipulate. Just as policymakers influence behavior by manipulating incentives and penalties through subsidies and fines, policymakers can influence behavior by manipulating the design of information technologies. Technology can operate as a means to regulate behavior analogous to law. This paper continues this line of inquiry by focusing on the role of default settings in software.

Default settings are pre-selected options chosen by the manufacturer or software developer. The software adopts these settings unless the user affirmatively chooses an alternative option. Defaults push users toward certain choices. For example, the most valuable part of Netscape was not its software, but its default setting for its home page. Because a large number of users never changed this default setting, Netscape’s home page had enormous popularity. Analysts touted the importance of this default home page (a top 10 web site at the time) when AOL purchased Netscape for \$4.2 billion [10].

Default settings are extremely relevant for policymakers today as there are a spate of proposals for proactively designing information technology to address issues such as crime, competition, free speech, privacy, protection of intellectual property, and the revitalization of democratic discourse. For each of these proposals, there are settings that must be set. These settings will inevitably favor certain options. By providing policymakers with guidance for setting software defaults, we can utilize defaults as a policy tool.

This paper contains several strands of thought in discussing defaults in software. At its base, this paper relies on an understanding of the interactions between technology and society. On top of that, this paper considers how software operates and is designed. This work is represented by our early comments on Human Computer Interaction and their insights on how defaults operate. Another strand is the work of legal scholars and behavioral economists who have studied defaults in a legal context. Our synthesis of these research strands provides an explanation for how defaults operate in software and how policymakers should set defaults in software. The general tone and focus of our work is policy based and leads us to provide

several examples applying our insights to contemporary policy concerns.

The article is organized as follows. The paper begins by defining settings in software including default settings. After an understanding of defaults, we move on to show how defaults have been traditionally analyzed in the law. This step is necessary, because we utilize the work of legal scholars in analyzing defaults in software. After establishing the concept of defaults, we review empirical data on the effectiveness of defaults. This research shows the importance and power of defaults. We then move on to explain why people defer to defaults. The final section of the paper focuses on how defaults should be set. Part of this normative section includes arguing the defaults are currently set incorrectly for two technologies that affect security and privacy.

2. DEFAULTS ARE ALONG A CONTINUUM OF COERCION

Software can be modified. The degree to which software can be modified can be seen along a continuum of coercion in Figure 1.

Fixed settings -----Default Settings-----Fully Customizable
“wired in” “pushing the user” “free choice”

Figure 1. Continuum of Coercion

The concept of defaults lies in the middle of this continuum and refers to a preselected option adopted by the code when no alternative is specified by the user. Defaults only exist when the user is able to change the default setting. A setting that the user is unable to change is a fixed aspect of the system (“wired in”), and is therefore not a default. Developers often used “wired in” settings for aspects of software that users do not need to modify.

A leading industry guideline for setting defaults comes from the Apple Human Interface Guidelines. It states:

The default button should be the button that represents the action that the user is most likely to perform if that action isn’t potentially dangerous... Do not use a default button if the most likely action is dangerous—for example, if it causes a loss of user data. When there is no default button, pressing Return or Enter has no effect, the user must explicitly click a button. This guideline protects users from accidentally damaging their work by pressing Return or Enter. You can consider using a safe default button, such as Cancel [1].

Besides industry guidelines, the setting of defaults is studied within the computer science sub field of Human-Computer Interaction (HCI) [8, 21]. An example of work within HCI is how Cranor, Guduru, and Arjula carefully considered the default settings in their design of the AT&T Privacy Bird, a P3P user agent [7]. The HCI research notes two considerations in setting defaults. First, default values can be set for novice users. The idea here is to protect novice users from adverse consequences. The second is that default settings should improve efficiency. This typically means setting a default value that most knowledgeable users prefer.

From a policy perspective, both existing rationales for setting defaults are far too vague. First, what is a novice user? It’s not

clear what defines a novice user, is it their knowledge, experience, education, or ability to use a computer? Why should we protect novice users? Secondly, efficiency is a slippery concept. Is the default setting most efficient for the software developers, expert users, or novices? Efficiency also assumes that it is possible to determine and calculate the costs and benefits of a default setting. However, many default settings have vague values such as privacy or externalities such as security that are difficult to calculate. While these rationales are undoubtedly useful to developers, they provide an insufficient basis for setting defaults from a policy perspective.

3. DEFAULTS IN LAW

Legal scholars have analyzed the role of default rules within the law. Their concept of default rules is analogous to the concept of defaults in software. For example, consider Barnett’s discussion about the default rule approach in the context of contract law:

The default rule approach analogizes the way that contract law fills gaps in the expressed consent of contracting parties to the way that word-processing programs set our margins for us in the absence of our expressly setting them for ourselves. A word-processing program that required us to set every variable needed to write a page of text would be more trouble than it was worth. Instead, all word-processing programs provide default settings for such variables as margins, type fonts, and line spacing and leave it to the user to change any of these default settings to better suit his or her purposes [3, p. 824].

For Barnett, the default rule approach refers to how certain obligations and responsibility are placed on the parties for a contract unless the contract specifies otherwise. This is similar to how a default in software places certain obligations or limitations on the users if they do not change it. The analogy with software improves because legal scholars have recognized that there are some rules that parties cannot change by contract. For example, Ayres and Gertner note that while the warranty of merchantability is a default rule that parties can waive, the duty to act in good faith is an immutable part of the contract that cannot be waived [2]. This notion of immutable rules is analogous to how rules may be wired into technologies.

Throughout this paper, our analysis of defaults is heavily influenced by the literature on defaults in law. However, we recognize there are two key differences between defaults in the law and in software. First, default rules in law are created by the state in the judicial or legislative branch. In contrast, defaults in code are typically created by software developers, which are striving to sell a product. Second, default rules are enforced quite differently. Parties seeking enforcement of a contract or damages for a breach of a contract go through a judicial process [24]. Part of the judicial process allows courts to refuse to enforce contracts that are unconscionable. There is no analogy to this enforcement or lack of enforcement in software. In software, enforcement is automatic and nonreviewable [9]. So while this paper relies on insights from the legal literature, there are differences between defaults in law and in software.

4. DEFAULTS ARE IMPORTANT AND EVERYWHERE

The malleability of software means that developers can add, remove, or change default settings. A typical program has tens (and up to hundreds) of defaults that are set by the developer. These defaults may be default values, which refer to strings, numbers, or bits that are held in a particular field for input screens or forms. Other defaults include default settings, which are values, options, and choices that are stored and referenced by an application. Finally, default actions are courses of actions that are presented to a user interactively. These defaults often come in the form of alert or confirmation boxes. In this paper, we use the term default or defaulting settings to refer to all three meanings of defaults in software.

Legal scholars have shown how default settings in law affect a wide variety of our everyday life from contracts [24] to labor and employment law [27]. Default settings in software also affect a variety of fundamental society policy issues. To illustrate this, we examine the defaults in a popular file sharing program known as Limewire.

Our examination of Limewire found several default settings that promote filesharing. This is not surprising. However, there are a number of other default settings affecting a variety of fundamental societal concerns. First, a default setting in Limewire sets the upload bandwidth default to 100%. This setting promotes using all available bandwidth for file sharing. Another default settings sets the program to automatically connect to the network when the application starts up. This ensures file sharing starts immediately. A third default setting treats users with fast computers and internet connections as an “ultrapeer.” An “ultrapeer” helps other users download faster, but inflicts a greater load on the user’s computer. All three of these default settings were used to promote filesharing. However, these are not the only defaults in Limewire. There are default settings for filtering search results by specific words, adult content, or file types. This setting affects free speech. Other default settings define the community of file sharers. Limewire has a default setting to only share files with people who are sharing files. Users can set the minimum number of files an uploader has to share. This feature defines the community’s boundaries. It can exclude “freeloaders” or people sharing only a few files. Limewire sets the default to 1 and thus effectively allows everyone to share files. Finally, there is a default affecting social communication. This default setting is used to turn the chat feature on or off.

The Limewire example shows how defaults can affect a wide variety of issues. As a matter of policy, defaults are good for a number of reasons. First, defaults provide users with agency. Users have a choice in the matter. They can go with the default option or choose another setting. Second, a default setting guides the user by providing a recommendation. However, there maybe situations where users do not need or should not have options. We discuss this in more detail later, but the key point is sometimes we do not want to give a user choices.

5. THE POWER OF DEFAULTS

A crucial issue for policymakers seeking to exploit defaults is whether defaults affect people’s usage. Are people sufficiently capable that they will configure software to their own needs and

disregard default settings? Or are people somehow swayed by default settings? This section reviews several studies in various contexts including 401(k) plans, organ donation, opt-in versus opt-out checkboxes, and wireless security.

Madrian and Shea found a significant shift in the saving behavior of individuals after a default was changed [17]. Initially, the default was set so that employees were not automatically enrolled in a 401(k) savings plan. The employer later changed this setting, so the new default setting was that employees would be automatically enrolled. Consequently, this new default resulted in an increase in participation from 37% to 86%!

In the case of organ donation, there are two possible defaults, either you are presumed to have consented to organ donation or a person must explicitly consent to donation. Johnson and Goldstein analyzed the role of default settings by looking at cadaveric donations in several countries [11]. They found that a strong effect of the default. When donation is the default, there is 16% increase in donation.

Bellman, Johnson, and Lohse examined the role of default settings in online checkboxes for opting-in or opting-out of certain practices [4]. These checkboxes are typically used for privacy settings, junk e-mail settings, and for a variety of other simple questions in online forms. They found the participation rate changed from 60% to 89% by changing the default from not participate to participate.

In the area of software, there are several anecdotal examples of the power of defaults. The first is Microsoft’s decision to change the default setting for the firewall in its operating system to increase security. The firewall is now turned on by default in Windows XP Service Pack 2. Microsoft tacitly understood that users defer to defaults and, therefore, security could be improved by changing this default setting. A second example is the fight over defaults for media players in Microsoft’s antitrust trial. RealNetworks offered a competing media player for Windows. It argued that PC manufacturers were not allowed to make any player other than Windows Media Player the default player [19]. Even if a user chose RealNetworks media player as the default player, Windows XP favored its own media player in certain situations [5].

Shah and Sandvig’s analysis of Wi-Fi access points examined the role of defaults in software [26]. Their research used a naturalistic methodology to examine how people configure their Wi-Fi access points (APs). They found defaults played a significant factor in how people used their APs. When a manufacturer set a default setting, this produced a compliance of 96-99%. However, when users were urged to change these default settings, only 28-57% of users acted to change the default settings. Moreover, about half of all users never changed any default settings on their APs. These results show how powerful defaults can be in software.

The explanation for this deference to defaults is not that people are uneducated or uninformed. In the examples above, with the exception of firewalls and wireless security, people understood the implications of deferring or changing a default setting. This is an important point, because it suggests that education alone does not account for why people defer to defaults. Instead, there are other factors that push people to defer to default settings. These factors can and have been harnessed by policymakers to influence how people act. For example, Thaler and Benartzi crafted a

savings program that took advantage of people's deference to defaults [29]. The result was increased savings by individuals.

6. EXPLAINING WHY DEFAULTS WORK

Once defaults are recognized as powerful, the next issue is to explain why people are swayed by default settings. Our synthesis of the legal literature, cognitive psychology, and our empirical data led us to four general explanations for why people defer to default settings. First, people don't understand what the default setting stands for. This is an issue of bounded rationality. Second, users may be subject to a cognitive bias that leads them to defer to the default setting. Third, even if users understand the default setting, they may be reluctant to change the default setting because they believe the default setting carries information about what is reasonable. Fourth, even if users want to change the default setting, they may lack the technical skills to actually change the default setting. The rest of the section discusses each of these explanations in greater detail.

The first reason is bounded rationality. People do not change defaults because they are uninformed. If a person does not know about the possibility of changing the option or the ramifications of each choice, a default setting is equivalent to a fixed setting. A good example of how people defer to defaults because they are uninformed or misinformed is the cookies technology. Cookies allow web sites to maintain information on their visitors, which raises privacy concerns.

A Pew Internet & American Life Project study from August 2000 examined online privacy concerns. Pew found that 84% of Internet users in the US were concerned about businesses and strangers getting their personal data online, but 56% did not know about cookies. More notably, 10% said they took steps to block cookies from their PCs. So while people were concerned about their online privacy, they were unaware of the most significant technology that affects online privacy. Not surprisingly, many people deferred to the default setting in their web browsers and accepted cookies. A study by Web Side Story found the cookie rejection rate was less than 1% [13]. This low rejection rate was used as evidence that most Internet users were not concerned about cookies. Several years later, JupiterResearch found that nearly 60% of users delete cookies, with 39% doing so monthly [30]. This is a significant change! Was it because people in 2004 were now informed about cookies and changing the defaults in their web browsers? There is little evidence that users are more sophisticated in using cookie management features in web browsers. For instance, a 2005 survey found that 42% of respondents agreed with patently false statements such as, "internet cookies make my computer susceptible to viruses" and "internet cookies make my computer unsafe for personal information." Another 30% admitted that they know nothing about internet cookies. Instead, the change in cookie behavior is largely attributable to the epidemic of spyware. Problems with spyware have fostered the growth of anti-spyware programs, which often delete cookies as part of their privacy protection. In the end, it appears the change in cookie management is because the default value in the web browser was trumped by another default value in the anti-spyware program.

A second reason people are averse to changing defaults is due to cognitive biases, including the status quo bias, omission bias, and

endowment effect. The status quo bias leads people to favor the status quo over a change. Samuelson and Zeckhauser explain the status quo bias as favoring inertia or having an anchoring effect [23]. The explanation is that individuals place greater value on the current state, and thus believe they are losing more if they make a change [12].¹

Another explanation for the status quo bias is described by the omission bias. The emphasis here is not on the current state, but on the fact that people often judge actions to be worse than omissions [22]. The omission bias suggests that individuals prefer to be hurt because some action was not taken, rather than equally hurt because some action was taken. In the realm of software, this bias suggests people will avoid changing a setting, because they fear "breaking" the computer.

The status quo and omission biases provide reasonable explanations for why people defer to defaults. To illustrate the differences between these explanations, consider a security setting for a firewall in a computer operating system. When a firewall is turned on, it provides the user with increased protection. Either bias could come into play in determining whether or not a user turns on the firewall, when the default is set for the firewall to be off. For example, a user knows that the firewall will protect her computer from certain hackers, but may be nervous about enabling the firewall because she is afraid it may "break" the computer. The status quo bias is suggesting that the current state (a working computer) is a safe state, and that leaving that state could result in a loss. Furthermore, the user is choosing to accept a possible harm due to omission versus a possible harm due to commission (turning on the firewall could lead the computer to malfunction). Therefore the omission bias comes into play.

Another reformulation of the status quo bias is the endowment effect. The endowment effect refers to how people value settings more when the default initially favors them, than when the default is set to favor another party. Empirical research has shown the endowment effect occurring as people often demand much more to give up something than they would be willing to pay to acquire it [15]. As we point out in later in the discussion section, the endowment effect raises the cost and, therefore the, difficulty in switching a default setting. Consequently, policymakers need to ensure the initial default settings are correct.

The third reason that people defer to defaults is known as a legitimating effect [27]. This arises because people believe defaults convey information on how people should act. Defaults are assumed to be reasonable, ordinary, and sensible practices. As a result, people can be resistant to changing a default setting. This assumption about defaults is not surprising, because under product liability law manufacturers have a duty to warn for dangerous products and a duty to "design out" dangers in a product. Consequently, when people use software, they assume that defaults are reasonable and sensible.

The fourth reason people do not change defaults is their lack of technical sophistication. In these cases, people know they ought to change the default, but can not figure out how. A crucial factor affecting this is the usability of software. Two examples that

¹ For a more complete overview of loss aversion, refer to the research of Tversky and Kahneman.

highlight this problem with software are security and pop-up advertising.

People are very concerned about security. An examination of software sales finds that security software is widely purchased by consumers. Two of the four best selling software titles in 2003 were system utilities and security products. Consequently, you would assume these informed and motivated individuals would have secure computer systems. However, in-home studies of computers have found considerable security deficiencies [18]. One study found that 80% of computers have spyware installed, 67% do not have updated anti-virus software, and 67% of users don not have firewall protections. The best explanation for this gap between purchasing software and continued security deficiencies is that people are unable to properly configure security software.

Another similar example concerns the inability of people to avoid pop-up ads. Surveys show that 77% of Americans find that telemarketing calls are "always annoying" and 78% of Americans consider pop-up ads "very annoying." In response to these annoyances, over 60% of households have signed up for the FTC's Do Not Call registry. In contrast, only about 25% of people have installed blocking software for pop-up ads. This discrepancy between the telephone and Internet is best explained by the technical difficulty of finding, installing, and configuring pop-up ad blockers as compared to signing up for the FTC's list.

7. HOW SHOULD POLICYMAKERS SET DEFAULTS

After acknowledging that defaults are powerful, the next issue is how policymakers can manipulate default settings. This section focuses on how policymakers ought to set defaults. After considering the threshold question of whether there should be a default setting, the section argues that defaults are generally good and that policymakers should allow parties to set default as they see fit. However, there are two circumstances when policymakers should intervene to switch default settings. They are when users are not properly informed and when significant externalities exist.

The first issue is whether there should be a default setting or a wired in setting. A wired in setting is analogous to the nonwaivable rights given to individuals, such as safety regulations and the right to family leave.² Sunstein notes several factors that policymakers should consider in analyzing whether a default setting is appropriate [28]. The first concerns whether users have informed preferences. If they know little about the setting, they are not likely to change it, and vice versa. It makes more sense to include a setting that people understand. The second issue focuses on usability and whether the mapping from options to preferences is transparent. In the case of software, this requires an easy to use interface that allows users to configure the software according to their preferences. The third issue focuses on how much preferences vary across individuals. If there is little or no variation in society it hardly makes sense to create a default setting as opposed to a wired in setting. The final issue is whether users

² Occupational Safety and Health Act, 29 U.S.C. §§ 651-678 (1994) (nonwaivable right to certain safety regulations); Family and Medical Leave Act, 29 U.S.C. §§ 2601-2654 (1994) (nonwaivable right to family leave).

value having a default setting. This can be determined by examining marketing materials, software reviews, and comments from users. If there is little concern over the default setting, it becomes reasonable for designers to opt for a wired in setting.

7.1 Defaults as the "Would Have Wanted Standard"

Behavior economists have analyzed how defaults should be set [27]. Much of this analysis has focused on defaults associated with law and social policy, specifically contracts, but this reasoning can be extended to software. The starting point is the Coase theorem, which holds that a default rule does not matter if there are no transaction costs [6]. This is because the parties will bargain to a common result that is efficient. Under this analysis, regulators do not need to be concerned with defaults in software, assuming there are no transaction costs. However, there are two problems with this formulation: transaction costs and cognitive biases.

There are always transaction costs in setting defaults. The general approach of legal scholars is that defaults should be set to minimize transactions costs. Posner argues that default rules should "economize on transaction costs by supplying standard contract terms that the parties would have otherwise have to adopt by express agreement" [20]. The idea here is that the default settings should be what the parties would have bargained for if the costs of negotiating were sufficiently low. This approach is known as "would have wanted" standard and is the general approach for setting defaults in contract law. This standard is also a good starting point for setting defaults in software. Let the parties decide what they want software to accomplish and then let the developers decide what options to build into software. In following this approach, developers would likely follow the common sense principles of HCI in protecting novices and enhancing efficiency.

There are two occasions when the "would have wanted" standard is not the optimal setting. In these cases policymakers may need to intervene. Before any intervention, it should be established that the default setting materially affects a fundamental societal concern. While it is not in society's interest for government to set the default font for a word processor, it is in society's interest to make sure fundamental societal values are protected.

7.2 Problem of Information

There are occasions when parties you would expect to change the default settings are not changing them. In this situation it is necessary to examine why the default setting is not being changed. For example, if defaults relating to accessibility are not widely changed, this should not raise a red flag, unless the disabled are also not able to change these default settings. If the disabled are not changing them, then there is an informational problem that is leading them to defer to the default setting.

If both parties are not fully informed, rational, and capable of changing the default settings, then the default should be what the parties "would have NOT wanted." The idea here is that this setting will force the parties to communicate and share information in order to change the setting to what the parties "would have wanted." In contract law, this is known as a penalty default and is used to encourage disclosure between the parties

[2]. A classic example of a penalty default is that courts assume a default value of zero for the quantity of a contract. The value of zero is clearly not what the parties would have wanted, since they were bargaining for an exchange of goods. However, this penalty default serves to penalize the parties if they do not explicitly change the default.

Penalty defaults are best used in situations where parties are not equally informed. In the case of software, this can mean uninformed, misinformed, or lacking technical sophistication. In everyday practice, this suggests that societally significant defaults should be set to protect the less informed party. This would force software developers to inform and communicate with users, when they want users to perform advanced actions that may have adverse consequences. In addition, it encourages developers to ensure that defaults can be changed with a minimal degree of technical sophistication.

An example where a penalty default is appropriate is the setting for cookies in web browsers. Cookies are a technology that allows web sites to maintain information on their users. Cookies are not well understood by most people. A penalty default would set the default to reject cookies. If web browsers and web sites want people to use cookies, then would have to explain to users what cookies are and how to turn them on. Unfortunately, the defaults found in web browsers nowadays are set to accept cookies. By changing this default, policymakers can harness the information forcing function of penalty defaults to improve the state of online privacy.

Penalty defaults are not appropriate in all circumstances, such as for settings that people readily understand. For example, if most people understand the concept and are capable of using software filtering technology to protect minors then a penalty default is unwarranted. In this case, policymakers should follow the would have wanted standard for setting defaults.

7.3 Externalities

A second reason for settings defaults at what the parties “would have NOT wanted” is to account for externalities. Settings in software can often affect third parties in a myriad of ways that are analogous to increasing the risk to an innocent passerby or pollution. In these situations, policymakers should consider the overall welfare of users and intervene to ensure a default value is set to reduce externalities. However, if the problem is grave enough, it may be necessary to change the setting from a default value to a wired in setting.

An example of where externalities are high is security. Most manufacturers would prefer not to enable all security functions, mainly because it leads to reduced functionality and increased supports costs. Moreover, most users know very little about security issues and cannot adequately bargain for their inclusion. However, this inaction costs everyone when computers are compromised. These costs could be reduced if security was enabled by default. Consequently, this suggests policymakers should ensure the default setting be set to enable security. Unfortunately, developers are still selling products that where the defaults are set to insecure values. The most egregious example is wireless access points that have default values for no security. Policymakers should force these developers to change their defaults to improve security and societal welfare.

7.4 Other Issues

The power of a default setting can be modified in two ways. The first is through changes in the user interface. This approach can affect how easy it is for a user to change the default. For example, increasing (or reducing) the prominence of a default setting can affect its use. Second, procedural constraints can make it more costly to change a default setting. The rationale for these constraints is to ensure users are acting voluntarily and are fully informed before they change a default setting. A simple example is an extra prompt that asks users whether they are really sure they want to change the default setting. These procedural constraints resolve problem of bounded rationality and bounded self-control. While a wide range of possible procedural constraints exist, they all serve to raise the cost of switching the default setting.

If modifications to the user interface and procedural constraints are not enough, then the situation may require a wired in setting versus a default setting. There are a variety of reasons, including safety and various externalities (e.g., radio interference, network congestion, or security), why users should not be able to change a setting. In these situations, a policymaker may seek a wired in setting, however, this is a serious decision, because it removes the flexibility away from the user.

In general, more options are better, because they allow users to reconfigure and use they software as they see fit. However, there are limitations to this rule. First, the more defaults present, the more likely users will be confused and intimidated by the number of choices. Second, there are practical limits to how many default settings designers can present in a useful manner without overloading the user interface. As an example consider the popular document format developed by Adobe called PDF. PDF viewers usually contain a handful of default settings. This simplicity allows everyday users to effectively use the software viewers. In contrast, the specialized software to create PDFs often contains tens to hundreds of default settings. The multitude of default settings allows creators fine-grain control over the production of PDFs. At the same time, the interface and use becomes daunting as designers have to find a useful way for organizing all these options. This problem of confusion and overloading places a practical limit on how many default options should be available to users.

8. DISCUSSION

Defaults affect a variety of fundamental societal concerns in software. This paper has noted how defaults in wireless routers affect security, defaults in cookies affect privacy, and pop-up advertising implicates free speech. These defaults are powerful and effectively de facto regulators. A variety of studies have shown how people defer to defaults regardless of how well they are informed.

The goal of this paper was not only to show that defaults are powerful, but also to provide policymakers with guidance on how to set default values. To this end, we first examined why people may defer to a default setting. We found that for a person to successfully change a default setting they need to be informed, overcome cognitive biases, be assured that its sensible to change the default, and finally have the technical skills. From this understanding we could then suggest how defaults ought to be set.

In general, policymakers should not intervene in default settings by relying on the “would have wanted” standard. This standard ensures the wishes of both parties are met in the design of defaults. However, there are two circumstances where policymakers may need to intervene and go against the settings agreed by users and developers. The first circumstance typically arises when users lack the knowledge and ability to change an important default setting. In these cases, policymakers ought to use penalty defaults to shift the burden of the default to the developer. This penalty default setting provides an information forcing function that serves to educate users and help them change default settings.

One scenario for a penalty default is for privacy issues. By setting a penalty default to protect a user’s information, this forces developers to notify and educate users before they have to share their personal information. While this approach is paternalistic, it still provides users with the freedom to choose as they wish. We suggest that in these rare situations when there is a fundamental societal concern at stake that people are uninformed, misinformed, or not technically sophisticated to change the default, then as a matter of public policy, people should be protected. If people want to give up that protection, then we should support well-informed individuals to make that decision. However, the default should be set to protect individuals.

The second circumstance where policymakers need to intervene involves default settings that cause harm to third parties. These externalities may need to be addressed by changing a default value. A good example of this is security. While it is in the interest of users and developers to make systems very open, this can have a negative externality because of costs from network congestion and spam. In this situation, policymakers have an interest in ensuring a default is either set to reduce externalities or insist that the default be replaced with a “wired in” setting to limit externalities.

A final consideration for policymakers is the initial default settings. As discussed earlier, the endowment effect is one reason that people defer to defaults. The endowment effect suggests the initial setting affects how defaults are valued. These valuations may make it very difficult for a later switch in a default setting. This effect means that policymakers need to carefully consider the initial default setting.

9. CONCLUSION

This paper shows how software can be used an alternative means of regulation. The first part of the paper provides empirical evidence on the importance of defaults and puts forth several explanations for why people defer to defaults. This part shows that altering default settings influences behavior. In the second part, we focused on how defaults should be set to promote societal welfare. This was illustrated by showing how security and privacy can be improved by changing default settings.

Our recommendation for setting defaults is that policymakers should not interfere in how defaults are set as a general rule. However, if a default affects a fundamental societal concern there are two circumstances when policymakers need to ensure the defaults are set to maximize overall social welfare. The first is when users lack the knowledge and ability to change an important default setting. This may require the use of penalty defaults that essentially set the default to what developers would not want. An

example of where a penalty default can be justified is in the cookie settings in web browsers. The second is when there are externalities present. Policymakers may need to intervene and switch a default to reduce overall societal costs. An example of where a default should be switched based on externalities is the security settings in software.

A normative analysis regarding settings in software is unique. While many scholars have recognized the power of software, ours is unique in arguing from a generalized framework what the settings of software should be. We believe that as scholars further investigate and understand the impacts of digital government, they will conduct normative analyses for other software characteristics. After all, policymakers have little guidance for analyzing other governance characteristics of software, such as transparency and standards.

In sum, this paper shows how software can be harnessed to address online issues. To this end, we describe two situations in which defaults are currently wrong. By ensuring these defaults are changed, policymakers can enhance security and privacy. We believe that the manipulation of software to enhance social welfare is a powerful tool and a useful complement to traditional legal methods.

10. REFERENCES

- [1] Apple Computer Inc. *Apple Human Interface Guidelines*, 2005.
- [2] Ayres, I. and Gertner, R. Filling the Gaps in Incomplete Contracts: An Economic Theory of Default Rules. *Yale Law Journal*, 99 (1989), 97-130.
- [3] Barnett, R.E. The Sound of Silence: Default Rules and Contractual Consent. *Virginia Law Review*, 78 (1992), 821-911.
- [4] Bellman, S., Johnson, E.J. and Lohse, G.L. To Opt-In or Opt-Out? It Depends on the Question. *Communications of the ACM*, 44, 2 (2001), 25-27.
- [5] CNN. Microsoft, RealNetworks Battle, 2002.
- [6] Coase, R. The Problem of Social Cost. *Journal of Law and Economics*, 4 (1960), 1-44.
- [7] Cranor, L., Guduru, P. and Arjula, M. User Interfaces for Privacy Agents. *ACM Transactions on Computer-Human Interaction* (Forthcoming).
- [8] Fowler, S.L. and Stanwick, V.R. *GUI Style Guide*. AP Professional, Boston, MA, 1995.
- [9] Grimmelmann, J. Regulation by Software. *Yale Law Journal*, 114 (2005), 1719-1758.
- [10] Herbert, D. Netscape in Talks with AOL, 1998.
- [11] Johnson, E.J. and Goldstein, D. Do Defaults Save Lives? *Science*, 302 (2003), 1338-1339.
- [12] Kahneman, D., Knetsch, J.L. and Thaler, R.H. The Endowment Effect, Loss Aversion, and Status Quo Bias. *Journal of Economic Perspectives*, 5, 1 (1991), 193-206.
- [13] Kelsey, D. Almost No One Rejects Cookies *Newsbytes*, 2001.
- [14] Kesan, J.P. and Shah, R.C. Shaping Code. *Harvard Journal of Law & Technology*, 18, 2 (2005), 319-399.
- [15] Korobkin, R. Endowment Effect and Legal Analysis. *Northwestern University Law Review*, 97 (2003), 1227-1293.

- [16] Lessig, L. *Code and Other Laws of Cyberspace*. Basic Books, New York, 1999.
- [17] Madrian, B. and Shea, D.F. The Power of Suggestion: Inertia in 401(k) Participation and Savings Behavior. *Quarterly Journal of Economics*, 116 (2001), 1149-1525.
- [18] National Cyber Security Alliance and America Online. Online Safety Study, 2004.
- [19] Orlowski, A. Why Real Sued Microsoft, 2003.
- [20] Posner, R.A. *Economic Analysis of Law*. Aspen Law & Business, New York, 2003.
- [21] Preece, J., Rogers, Y., Sharp, H., Benyon, D., Holland, S. and Carey, T. *Human-Computer Interaction*. Addison-Wesley, Wokingham, UK, 1994.
- [22] Ritov, I. and Baron, J. Status-quo and Omission Biases. *Journal of Risk and Uncertainty*, 5, 1 (1992), 49-61.
- [23] Samuelson, W. and Zeckhauser, R. Status Quo Bias in Decision Making. *Journal of Risk and Uncertainty*, 1, 1 (1988), 7-59.
- [24] Schwartz, A. and Scott, R.E. Contract Theory and the Limits of Contract Law. *Yale Law Journal*, 113 (2003), 541-601.
- [25] Shah, R.C. and Kesan, J.P. Manipulating the Governance Characteristics of Code. *Info*, 5, 4 (2003), 3-9.
- [26] Shah, R.C. and Sandvig, C., Software Defaults as De Facto Regulation: The Case of Wireless Access Points. in *Telecommunications Policy Research Conference*, (Washington, DC, 2005).
- [27] Sunstein, C.R. Switching the Default Rule. *New York University Law Review*, 77 (2002), 106-134.
- [28] Sunstein, C.R. and Thaler, R.H. Libertarian Paternalism is Not an Oxymoron. *University of Chicago Law Review*, 70, 4 (2003), 1159-1202.
- [29] Thaler, R.H. and Benartzi, S. Save More Tomorrow: Using Behavioral Economics to Increase Employee Saving. *Journal of Political Economy*, 112, S1 (2004), S164-S167.
- [30] Wegert, T. The Web Cookie is Crumbling -- and Marketers Feel the Fallout *Globe and Mail*, 2005.

Experimental Application of Process Technology to the Creation and Adoption of Online Dispute Resolution

Ethan Katsh, Leon J. Osterweil,
Norman K. Sondheimer
University of Massachusetts Amherst
Amherst, MA 01003
1-413-545-4228
Katsh@legal.umass.edu
LJO@cs.umass.edu
Sondheimer@cs.umass.edu

ABSTRACT

We report on the development of formal models of alternative dispute resolution processes, the creation of an online dispute resolution system based on this model and initial experimental analysis of this system. Early results suggest that formalizing the negotiation process definition indeed leads to clearer understandings and a greater chance for effective automation.

Categories and Subject Descriptors

K.4.3 [Organizational Impacts]: Computer-Supported Collaborative Work

General Terms

Experimentation

Keywords

Online Dispute Resolution, Process Technology, Mediation

1. INTRODUCTION

The use of alternative dispute resolution (ADR) processes has grown rapidly over the last three decades. In the last few years, the field of online dispute resolution (ODR) has developed to enable mediation and arbitration to occur at a distance and to use computers to enhance and assist in the resolution of conflict. ODR has taken hold in e-commerce. Like many other areas of Information Technology, ODR has been slow to take hold in the Federal Government. We hypothesize that the key to adoption lies in effective change management processes. We contend that such processes can be based on merging powerful process definition and analysis approaches into participatory computer systems design methods. This note reports progress on a project undertaken by the University of Massachusetts Amherst and the National Mediation Board (NMB) to understand how ODR can improve efficiency, effectiveness, and fairness in Government dispute resolution and how ODR systems can gain acceptance.

Daniel Rainey

National Mediation Board
1301 K Street, N.W., Suite 250 East
Washington, D.C. 20005
1-202-692-5051

Rainey@nmb.gov

2. OUR APPROACH

We argue that technology can be looked at as a “fourth party,” an element in the dispute resolution process that can play various roles in consensus building, in decision making, and in the interaction between the parties in dispute and a third-party neutral [1]. We contend that ODR will offer increased access for public participation, more effective public policy processes, and new processes for collaborative problem solving. ODR provides asynchronous and/or real-time capabilities which can be leveraged with the ability to bring people together virtually. ODR offers new tools for multi-party stakeholder collaborations, facilitation, negotiation, and mediation. One of the most widely known processes for dispute resolution is Interest-Based Bargaining (IBB) [2]. NMB offers IBB mediation. We have based our initial ODR process on IBB.

At the core of the differences between commercial and government adoption of technology is the overriding need to establish cooperation between stakeholders to permit change. We hypothesize that we can enhance the adoption of digital government including ODR by building on process technology. We view dispute resolution as a complex process, whose clear, precise, and complete definition will pave the way for development of efficient, effective and fair ODR systems. We have applied our research on process languages and formalisms to the problem of defining these ODR processes [3]. This work forms a solid basis for studying what must be changed or added to meet the challenges of ODR for grievance mediation. To this end, we are developing methods to embed process technology in participatory design methods, such as Joint Application Development (JAD), to ease the acceptance of the resultant ODR systems. Together, we believe we can produce efficient, effective and fair methods of producing ODR systems that will be readily adopted.

3. MODELING THE NMB IBB GRIEVANCE MEDIATION PROCESS AND SUPPORTING IT ONLINE

The NMB, established by the 1934 amendments to the Railway Labor Act of 1926, is an independent agency that helps facilitate harmonious labor-management relations within the nation's railroads and airlines. NMB IBB programs provide an integrated dispute resolution process including face-to-face mediation processes typical of many in person mediation processes.

While casual observers may be skeptical that process formalism can facilitate the informal process of negotiation, skilled negotiators, such as those at the US National Mediation Board (NMB), have long understood that careful adherence to predefined process restrictions can do much to facilitate the process of bringing disputants to agreement. Over a period of many years and decades, much has been learned and documented about how to discipline negotiation processes. These disciplined processes are currently passed from person to person through instruction and training, and indeed the understanding of the nature of such processes grows accordingly over time.

We are using a rigorous process definition language to develop and exhibit the first rigorous definition of the IBB process used by NMB. We have been using the Little-JIL process formalism as the basis for this definition. Little-JIL is one of a family of process definition languages that are defined through formal semantics that create the possibility of process definitions that demonstrate precision and rigor. Over the last year we have been trained by NMB in their processes, interviewed NMB mediators on their standard practices and recorded their recollections of mediations they have conducted.

Using the process model as a specification, we have first concentrated on a prototype ODR tool for supporting a brainstorming process that is at the core of the NMB's grievance mediation method. The mapping from manual process to online tool has been relatively straight-forward. In fact, the willingness of NMB to support this research stems in part from the complexity of commercially available tools. The tools support many different types of mediation, not just the one NMB has settled on. The tool is described in more detail in an accompanying note [4]

4. EVALUATION

The ODR system has been demonstrated on numerous occasions and subjected to several rounds of experimental evaluation. Three University classes on the University of Massachusetts Amherst campus have simulated work as parties in NMB training cases. Half of each class was assigned to act as a team representing each party. One of the three, an online University class with members across the U.S., has used the ODR system to run an extended asynchronous mediation case study. All of these had a professional mediator running the session. NMB professional mediators have used this ODR system in simulations based upon cases. Here one set of mediators was assigned to act as the mediator on the case.

At a high level, we have found that participants have little trouble adopting the online tool, effectively generating many ideas and moving fairly directly towards solutions. At the same time, the volume of text generated has presented a challenge to effectiveness. Participants and instructors noted that the anonymity of posts lowered inhibitions for questionable contributions. In addition, inhibitions for creative engagement in brainstorming were also lowered – and this enhanced the quality and content of the posts.

Response from the mediators is especially promising. They have in the past expressed dissatisfaction with the complexity of their

existing ODR tool. Our hypothesis has been that involving them in the design of the new tool will aid change. Having several of the subjects as models for the mediation process, we would be disappointed if they did not approve of the tool. We were not. Here are quotes from three mediators: “I am amazed that software is so far along ... (sic) & that it is so user friendly. My enthusiasm is directly proportional to the ease of computing.”; “Great capabilities w/ (the) software.”; and “Already an easy-to-use system in its prototype phase.”

At the same time, the mediators in a JAD session, as well as the students through written surveys, have suggested many improvements on the prototype. We have a list of over 50 changes and enhancements. Suggestions for modifications range from screen layout to text edit functions. Suggestions for new functionality include the possibility of concurrent discussions and support for reorganizing lists of ideas. The majority of suggestions have come from the mediators. We take this to indicate that the careful analysis of their process has led them to a clearer understanding of the possibilities of computer support.

5. NEXT STEPS

We are currently evaluating the trade offs of expanding the functionality of the system. We are developing tools to directly connect the process model to the prototype so stakeholders can see the implications of changes in the process directly. We continue to explore ways to realize the promise of the computer as a Fourth Party to fully “assist in identifying and evaluating interests, options and solutions”.

6. ACKNOWLEDGMENTS

This material is based upon work supported by funds from the National Science Foundation under Grant No. IIS-0429297, as well as, funds from the National Mediation Board. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of the National Science Foundation or the National Mediation Board

7. REFERENCES

- [1] Katsh, E. and Rifkin, J., *Online Dispute Resolution: Resolving Disputes in Cyberspace*. San Francisco: Jossey-Bass, 2001.
- [2] Fisher, R. and Ury, W., *Getting to Yes: Negotiating Agreement Without Giving In*. New York: Penguin Books, 1983.
- [3] Cass, A.G., Lerner, B.S., McCall, E.K., Osterweil, L.J., Sutton Jr., S.M., and Wise, A. *Little-JIL/Juliette: A Process Definition Language and Interpreter*. in *International Conference on Software Engineering*. Limerick, Ireland, 2000.
- [4] Clarke, L., Gaitenby, A., Gyllstrom, D., Katsh, E., Marzilli, M., Osterweil, L.J., Rainey, D., Sondheimer, N.K., Wing, L., and Wise, A. *A Process-Driven Tool to Support Online Dispute Resolution*. in *DG.o2006: The 7th Annual International Conference on Digital Government Research*. San Diego, 2006.

SESSION 7B

DIGITAL DOCUMENT PRESERVATION AND ARCHIVING

Moderator

Laura Steinberg, George Washington University, USA

Titles and Authors

Building a state government digital preservation community: Lessons on interorganizational collaboration
Kwon, Hyuckbin; Pardo, Theresa A.; Burke, G. Brian

Robust Technologies for Automated Ingestion and Long-Term Preservation of Digital Information
JaJa, Joseph

Building a Demonstration Prototype for the Preservation of Large-Scale Multimedia Collections
Rajasekar, Arcot; Berman, Francine; Burstan, Lynn; Kreisler, Harry; Schottlaender, Brian;
Moore, Reagan; Marciano, Richard; Hou, Chien-Yi; Anderson, Steve; McEwen, Mellisa;
Bornheimer, Bee; DeClerck, Luc; Westbrook, Brad; Hutt, Arwen; Kozbial, Ardys;
Fryman, Chris; Chu, Vivian

DIGARCH Project Highlights: Multi-Institutional Testbed for Scalable Digital Archiving
Miller, Stephen P.; Detrick, Robert S.; Helly, John

Building a state government digital preservation community: Lessons on interorganizational collaboration

Hyuckbin Kwon, Theresa A. Pardo, and G. Brian Burke

Center for Technology in Government
University at Albany, State University of New York
187 Wolf Road, Suite 301
Albany, NY 12205
{hkwon, tpardo, bburke}@ctg.albany.edu

ABSTRACT

As a part of the National Digital Information Infrastructure and Preservation Program (NDIIPP), the Library of Congress sponsored a series of collaborative workshops between April and May 2005 to help state governments identify their needs and priorities for digital preservation. During these workshops, state and territory representatives showed strong interest in fostering partnership efforts and collaborative strategies toward preserving state government digital information. Based on the findings of the workshops and previous efforts on digital preservation, this paper discusses the challenges and opportunities regarding interorganizational collaboration and community building for digital preservation of state government information.

Categories and Subject Descriptors

K.6.0 [Management of Computing and Information Systems]

General Terms

Management, Human Factors, Standardization

Keywords

Digital Preservation, Interorganizational collaboration

1. INTRODUCTION

The rapid development of information technology has dramatically changed the way information is created, stored, and used in the public and private sectors in the United States. At the state government level, vast amounts of information is created in electronic form, including land data, school records, official publications and court records. For instance, a recent study [8] reports that over 50% of North Carolina state government publications are produced and disseminated in digital format only. Although the digitization

of government information can promote efficiency, searchability and accessibility, it involves difficult challenges as well; the long-term preservation of electronic records is one of them. Much of electronic government information is of permanent legal, legislative, or cultural value, yet is at significant risk of loss because of fragile media, technological obsolescence, and other difficulties. As a 2003 American Association of Law Libraries study concludes, however, the need to preserve electronic government information is "yet unmet in any comprehensive manner either at the federal, state or local level." [9]

In order to address these issues, Congress enacted the National Digital Information Infrastructure and Preservation Program (NDIIPP) legislation in December 2000. The legislation charges the Librarian of Congress to lead a nationwide planning effort for the long-term preservation of digital content, as well as to capture current digital content that is at risk of disappearing. [13] As a part of the NDIIPP, the Library of Congress (LC) aims to include state governmental entities (state libraries, archives, and other state agencies) in the national network to preserve "born digital" state and local government information that is both significant and is at risk of loss. The Center for Technology in Government (CTG), a digital government research center at the University at Albany, has been working with the LC since September 2004. The main responsibility of CTG is to develop a capability assessment and planning toolkit [11] to support the preservation efforts of state governments.

Between April and May 2005, LC sponsored three workshops to help states identify their needs and priorities for digital preservation. CTG played a key role in planning, facilitating, and analyzing the results of the workshops. This paper reports the findings of the workshops and discusses the challenges and opportunities regarding interorganizational collaboration and community building for digital preservation of state government information.

2. THE LIBRARY OF CONGRESS CONSULTATION WITH STATES WORKSHOPS

2.1 Purpose and Audience

Beginning in March of 2005, LC invited U.S. states and territories to form collaborative arrangements and develop strategies for preservation of significant state and local

government information in digital form. The invitations were sent to the heads of state libraries and state archives and territorial equivalents. LC requested that each state library and archives consult between themselves and also as appropriate with other stakeholder entities in their state to determine the composition of the best team to participate in one of the three workshops. In the invitation, the Library indicated that it was strongly interested in active collaborations within and between states to address a shared approach to digital preservation. The Library stated that, ideally, this approach draws on an association among various entities with a stake in the long-term management and preservation of government digital information in each state, such as the state library, archives, records management organization, county clerks and other agency information custodians, and chief information officer (or information resource executive).

The purpose of the workshops was to collect facts, perspectives, and recommendations regarding digital preservation of state government information from librarians, archivists, records managers, information technologists, and other professionals representing U.S. states and territories. LC, in collaboration with the Center for Technology in Government, used the workshops to work with the state and territorial participants to collect this information through a series of large and small group facilitated discussions and exercises.

The three one-day workshops were held on April 27th, May 11th, and May 25th. The first and third workshops were held in Washington, DC, and the second one was held in Baltimore, Maryland. Three separate dates were selected in order to facilitate participation from all states, territories, and the District of Columbia. All 50 states, the District of Columbia, and three territories sent representatives to one of the three Spring workshops. Across the three workshops, 67 librarians, 53 archivists, 13 records managers, and 20 IT professionals were in attendance. While it was up to the individual participants and their other state or territory representatives to select the workshop date, each of the three workshops had a geographically diverse mix of states in attendance. Each workshop had between 14 to 19 states and at least one territory represented. Also in attendance at each of the three workshops were a small group of observers from other federal agencies and professional associations interested in digital preservation, including National Archives and Records Administration (NARA), Institute of Museum and Library Services (IMLS), Government Printing Office (GPO), Council on Library and Information Resources (CLIR), National Historical Publications and Records Commission (NHPRC), and Council of State Historical Records Coordinators (COSHRC).

2.2. Workshop summary

Each workshop was structured to include presentations on NDIIPP and large and small group-facilitated discussions and exercises involving all of the state and territory representatives. A round robin large-group discussion focused on top concerns relating to digital preservation, major success stories, and areas of interest to discuss with

other states. And small group breakout sessions, facilitated by CTG and LC staff, focused on three basic issues of concern to LC about preservation of state government digital information in the states and territories. For all three workshops, state and territory teams were assigned to one of four small groups. Each of the four small groups was comprised of between 4-6 states. Each small group spent between 45 to 60 minutes working on exercises and engaging in facilitated discussions focused on the following three questions:

1. What kinds of digital content are at-risk and what are the priorities for preservation?
2. How can states extend or build partnership networks?
3. What preservation-related roles do states and the Library need to fill?

2.2.1. At-risk state government digital information

As shown in Table 1, the categories of information that are considered most at risk by the state participants were government records, databases, digital publications, Web sites, and e-mails.¹ There were also informative discussions on issues concerning particular types of content, such as voluminous and dynamic characteristics of Web sites and e-mails and migration concerns on legacy documents and obsolete formats.

2.2.2. Preservation Partnerships

The workshop participants identified many existing networks that currently support partnerships for digital preservation. The networks identified in all three workshops are:

- Within states: municipal and local associations, task forces, GIS community
- Between states: National Association of Government Archives and Records Administrators (NAGARA), Online Computer Library Center (OCLC)
- Between states and private sector: OCLC
- Between states and federal government: NHPRC, National Endowment for the Humanities (NEH), IMLS, LC/NDIIPP, NARA, GPO

Also, the participants in all workshops regarded information sharing and education as a means to leverage partnerships, and competing priorities, lack of funding, lack of knowledge, and different perspective of IT people as barriers to partnerships.

¹ Note that some categories are not mutually exclusive. Since the characteristics of workshop discussions was close to that of brainstorming sessions, the classification of categories was not done in a very rigorous manner. For example, Web contents fall into government publication (by the nature of content itself) and Web sites (by the media). For more detailed data from the workshops, including the number of votes, see [14].

Table 1. At-risk state government digital information, in order of importance as voted by workshop participants

At-risk digital information	Examples
Records	born-digital official records, legal records, legislative records, property records, working documents, poorly scanned materials without hard copies
Databases	e-government transactional databases, GIS, fiscal databases, electronic filings, agency records in database format
Digital publications	government publications, Web-based publications, statistical reports, forms, information about state
Websites	Web contents of value, state government Web sites, agency Web sites, governors' Web site
Email	agency e-mail, public and private correspondence, links, instant messaging, official e-mail records, Public records in email format
Data sets	GIS, voter list, legacy data, data files
Audio & Video	multimedia, digital video and photos, digital recordings of legislative proceedings and public meetings, public broadcasting
State-wide elected officials and agency heads	governor's, Attorney General's, state legislature
Geographic information systems (GIS)	
Migration issues	legacy documents, legacy systems proprietary, obsolete formats
Internal Documentation	electronic source documents for subject files, developmental process behind documents
Document conversion	digital images
E-filings transactions	court records, vital records, deeds, wills
Restricted information	
Cultural heritage	history and culture, indigenous languages
Administrative metadata	
Maps	

2.2.3. State and LC Roles and Responsibilities in Support of Digital Preservation

Preservation-related roles and responsibilities for LC, in order of importance as voted by workshop participants, are as follows:

- Funding
- Best practices
- Coordination/facilitation/ Partnership
- Clearinghouse
- Standards
- Training/ Education
- Advocacy
- Archiving
- Promotion
- Direct services

Providing funding, developing best practices, and promoting collaboration/facilitation were LC's roles that received most votes in all three workshops. Training/education and development of standards were common items as well.

The roles for state governments, in order of importance as voted by workshop participants, are:

- Records selection/ Collection management
- Legislation/policy and Legal issues
- Access
- Communication/ Collaboration
- Funding
- Leadership/ Advocacy/ Education
- Strategic planning
- Setting priorities
- Creating infrastructure
- Guidance to employees
- Partnerships
- Standards
- Involving stakeholder
- Foundation
- Collecting and preserving its own records
- Implementation
- Demo projects
- Building the infrastructure without duplication
- Statewide digital initiative
- Technological tools

3. DISCUSSION

3.1. Interagency and Interprofessional Collaboration

The main actors in a digital repository system are producer (information provider), management (professional), and consumer (user) [2, 3]. The collaboration among these actors as well as within each class of actors is crucial for ensuring the preservation of and the long-term access to digital records. More specifically, collaborative efforts in digital preservation can bring the following benefits [15]:

- Access to a wider range of expertise
- Shared development costs
- Access to tools and systems that might otherwise be unavailable
- Shared learning opportunities
- Increased coverage of preserved materials
- Better planning to reduce wasted effort
- Encouragement for other influential stakeholders to take preservation seriously
- Shared influence on agreements with producers
- Shared influence on research and development of standards and practices
- Attraction of resources and other support for well-coordinated programs at a regional, national or sectoral level

For the successful digital preservation of state government information, an agency responsible for preservation, as the management of system, needs to leverage partnerships with various stakeholders such as private sector entities, other state governments, the federal government, local governments, other branches of state government, and other state government agencies. The following discussion focuses on interagency and interprofessional collaboration among librarians, archivists, records managers, and IT staff, which was one of salient issues in the workshops.

Most research in digital libraries so far has taken systemcentric approaches to address how the service will be provided and does not explore in detail the roles of and the relationships between different actors in the digital preservation community [2]. Particularly, the influence of different perspectives and behaviors of these actors on interactions between them in public sector was rarely examined. Although not specifically focused on long term preservation of digital information, there have been collaborative efforts between librarians, archivists, and information technologists for electronic records management in academic institutions. The Coalition for Networked Information (CNI) was formed in 1990 to bring together the content expertise of librarians and the networking expertise of information technologists. According to CNI's Working Together workshop report [7], the factors motivating collaboration include executive mandates, scarcity of financial resources, the interdependence of librarians and information technologists, the desire to consolidate overlapping functions, the need to incorporate the other professional group's perspectives into project design, while time and costs needed for partnerships, differences in organizational

culture, lack of respect for the other profession, and personality conflicts mitigate against successful partnerships. Also, McGovern and Samuels [11] emphasize the importance of collaboration between archivists and IT staff at colleges and universities. Such partnerships bring together archivists' knowledge on the value and context of records, identification and selection of content, and legal issues and information technologists' expertise on structure of records, networked environment, and technical issues. The authors contend that other professionals such as legal counsel, auditor, and financial officers also need to join this partnership for successful electronic records management.

Some academic studies in other areas such as health care and criminal justice contain detailed discussions on interprofessional and interdisciplinary collaboration. For example, Hall [5] explains the influence of different professional cultures on interprofessional teamwork. Professional culture, which includes values, beliefs, attitudes, customs, and behaviors, is established by means of education and socialization and remains obscure to other professions. Although different cultures pose challenges such as unfamiliar vocabulary, different approaches to problem solving, and a lack of common understanding of issues and values, they can lead to synergistic efficiency, creative solutions, and improved job satisfaction if properly leveraged.

Workshop attendees with different professional backgrounds expressed different concerns and interests regarding digital preservation [14]. Librarians tend to emphasize permanent public access and item-level description and control. On the other hand, the archival focus was on handling aggregates rather than items. IT staff were generally less concerned with information itself and were more interested in methods for information management and control, particularly system security. As for content types, librarians regarded electronic publications most important, while archivists and records managers were most concerned with the preservation of public records.

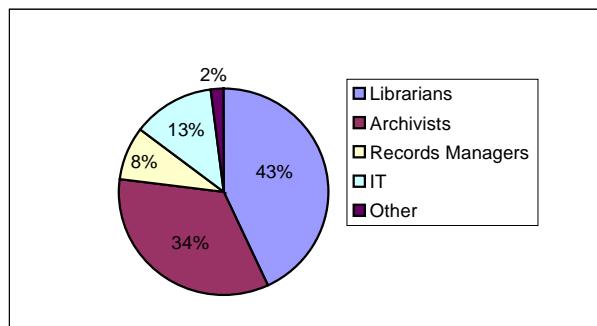


Figure 1. Summary of workshop participants

The contradiction between librarians/archivists and IT staff was particularly salient in the workshop discussions. The workshop participants, mostly composed of librarians, archivists, and record managers as shown in Figure 1, listed as barriers to successful partnerships different professional perspectives, backgrounds, and work cultures between

librarians/archivists and technologists, professional stereotypes, lack of bridging professionals, and IT staff's lack of knowledge on library networks, and suggested closer relationship between librarians/archivists and state CIOs, educating IT people on archivists and librarians' work and getting different professionals to talk together. The lack of shared language between archivists and information technologists leads to poor communication between the two professional groups. For example, for archivists the term *archives* is a noun which refers to a place where public records or other important historic documents are kept, or the records or documents that are so preserved. But for information technologists, *archive* is a verb meaning to transfer information to a storage location containing infrequently used files, for example, from disk to tape. [1]

Interagency settings in state governments pose more challenges to collaboration in digital preservation. In many cases, as stated by workshop participants, state libraries, archives, records management agencies, and IT departments have formed multiple silos and battle for their "turf". As a result, the communication and sharing of information across these agencies are hindered, and the collaboration becomes more difficult. Based on the results of their international case studies, Dawes and Prefontaine [4] assert the need for a formal institutional framework and relevant technology choice for successful interorganizational collaborations in the public sector. These themes appear consistent with the findings of the workshops in several ways. First, the institutional legitimacy for the digital preservation partnership began with a basis in law (the NDIIPP legislation) and was reinforced by the sponsorship of a recognized authority (LC). The state representatives showed a strong willingness to gather together on a regular basis and network with one another. The establishment of more formal partnership structures between states would facilitate more communication and secure the collaboration against political changes. Second, the choice of technology tools, especially metadata and preservation standards, was one of the main topics of the workshop discussions. Many attendees regarded the development and enforcement of national standards as one of the critical roles of LC. However, the findings of Dawes and Prefontaine imply that such tasks will be challenging ones, as the nature, cost, and cost distribution of the technology choice will have a significant influence on the participation and performance of this initiative. The fact that many agencies have interests in the metadata and preservation standards they have already chosen and are using is likely to further complicate this issue.

As shown in Table 2, UNESCO *Guidelines for the Preservation of Digital Heritage* provides four structural models of collaboration for digital preservation [15]. Among these models, the centralized distributed model appears to be most relevant for the digital preservation of state government information at this stage, since LC is capable of and willing to take responsibility as facilitator and coordinator. As the workshop findings regarding the

roles for LC and states suggest, LC can assist states to identify and preserve their own records by providing funding and coordinating standards setting processes. This way participants can benefit from economies of scale in infrastructure investments and diverse expertise and experiences.

3.2. Building a Digital Preservation Community

Wenger's theory of "community of practice" [16] provides useful insights on why and how the digital preservation community should be established. Communities of practice are "groups of people who share a concern, a set of problems, or a passion about a topic, and who deepen their knowledge and expertise in this area by interacting on an ongoing basis." [17] They operate as "social learning systems" where practitioners connect to solve problems, share ideas, set standards, build tools, and develop relationships with peers and stakeholders. Because they are inherently boundary-crossing entities, communities of practice are a particularly appropriate structural model for cross-agency and cross-sector collaborations.

The librarian community, the archival community, and the information technology community can be regarded as separate communities of practice in that they consist of self-selected members, aim to develop member's capabilities and exchange knowledge, and are held together by passion, commitment, and identification with the group's expertise. [18] According to Wenger [16], different communities of practice can be interconnected by boundary objects (reificative connection) and brokering (participative connection). First, the reificative connection is provided by shared artifacts, documents, tools, concepts, and other objects around which communities of practice can organize their interconnections. Second, participative connection is provided by people with multimembership who can introduce elements of one practice into another. The two are complementary in that boundary objects can overcome the physical limitation of participative connections, and brokering can solve the problem of ambiguity in reificative connections. When the connection between different communities of practice becomes established and provides an ongoing forum for mutual engagement, it can produce a new boundary practice, and ultimately a community of practice in its own right. Many communities of practice, including new scientific disciplines, have been established in this way.

The findings of the workshops indicate the need for connecting different communities and creating a new community of practice for digital preservation. First, the majority of participants demanded best practices and standards for digital preservation, which are reificative objects that can provide a means of coordinating different perspectives. Second, there was strong interest in meeting again to regularly revisit the issues facing digital preservation efforts. This is considered to be evidence of participatory connections across communities of practice.

Table 2. Structural models of collaboration for digital preservation

	Centralized distributed model	More equally distributed model	Very highly distributed collaboration	Standalone arrangements
Structure	<ul style="list-style-type: none"> • Consists of a partner that leads on policy, sets directions and provides most of the infrastructure, working with many others who have clearly specified but limited roles, such as identifying materials to be preserved and adding metadata, with limited responsibility for long-term maintenance 	<ul style="list-style-type: none"> • Consists of a number of partners with similar levels of commitment and responsibility 	<ul style="list-style-type: none"> • Consists of a large number of partners, each playing a very restricted role, perhaps limited to self-archiving 	
Strengths	<ul style="list-style-type: none"> • Cost sharing • Pool of ideas and perspectives • Economy of scale • Better controlled processes • Efficient decision making 	<ul style="list-style-type: none"> • Cost sharing • Pool of ideas and perspectives • Encourages shared level of ownership • No pressure of making decisions alone 	<ul style="list-style-type: none"> • Low costs for each partner • Useful starting point for a preservation program, raising awareness and allowing some steps to be taken 	<ul style="list-style-type: none"> • May contribute to later collaboration by allowing programs to develop expertise, strategies and systems before looking for suitable partners
Weaknesses	<ul style="list-style-type: none"> • May not encourage ownership of the program among the peripheral partners • May not be effective in encouraging transfer of skills from the central agency 	<ul style="list-style-type: none"> • May be difficult to establish effective leadership • Consultation and decision making may be time-consuming • Economy of scale may be lost 	<ul style="list-style-type: none"> • Unlikely to offer much reliability without a large investment in specifications, training and checking • May lead to high costs overall • May have trouble addressing long-term preservation issues in a coordinated way 	
Relevant areas	<ul style="list-style-type: none"> • Beginning programs seeking to collaborate with large, advanced programs • One program willing to take ongoing responsibility and a number of others who can help but are not sure about their long-term commitment 	<ul style="list-style-type: none"> • A number of players willing to share responsibility but none wanting to lead a program 	<ul style="list-style-type: none"> • A number of small sites capable of taking some limited responsibility, especially if there is one partner able to play a coordinating role • Materials for which preservation is desirable rather than essential 	<ul style="list-style-type: none"> • Programs operating in an environment where there are no suitable potential partners
Example	<ul style="list-style-type: none"> • A central records authority working with government business agencies, setting standards and providing guidance 	<ul style="list-style-type: none"> • A group of data archives that decide to agree on standards and share specifications for purchasing computer equipment 	<ul style="list-style-type: none"> • Networks of local community projects that decide that they will all keep their material for posterity 	<ul style="list-style-type: none"> • A small research facility decide that its data must be preserved and set up a modest program to document, back up and migrate its data, hoping to eventually find a program that will take responsibility for it

Note: Based on Webb, C. *Guidelines for the Preservation of Digital Heritage*. United Nations Educational, Scientific and Cultural Organization, Paris, Mar. 2003. 62-67.

The creation of a state government digital information preservation community would allow reconciling different perspectives of librarians, archivists, records managers, and IT staff, and utilizing their expertise for successful digital preservation. Snyder et al [12] illustrate examples of successful communities of practice in the federal government. Among these examples, the case of the e-regulation community appears particularly relevant to our discussion. The e-regulation community consists of professionals in IT and knowledge management, and records management from ten federal agencies and aims to develop an electronic compliance and records management system. The community, which started from an effort to share best practices with other agencies to meet statutory mandates, has promoted cross-agency collaboration and knowledge-sharing. Following these discussions on community of practice, a “state government digital information preservation community” could be structured as in Table 3, and incorporated into the national digital preservation network.

Table 3. State digital preservation community

Sponsor	The Library of Congress
Domain	Preserving the digital information of state governments
Members	Librarians, archivists, records managers, and CIOs and IT staff in state and territorial governments
Activities	Meetings, sharing best practices and project ideas, joint projects
Outcomes	Increased collaboration between states and within a state

4. CONCLUSIONS

One of the basic themes that emerged from 2005 Library of Congress Consultation with States Workshops was the need for collaboration among librarians, archivists, records managers, and CIOs and IT staff to preserve the digital information of state governments. The workshop findings show that the information professionals in state governments are willing to collaborate with one another, but face challenges such as different interests and professional culture, a lack of common understanding of issues and values, and language barriers. In order to reconcile different perspectives of information professionals and utilize their expertise, we make the following recommendations.

First, establish a “state government digital information preservation community” and incorporate it into the national digital preservation network. The community, composed of librarians, archivists, records managers, and CIOs and IT staff in state governments and supported by LC, could promote collaboration for digital preservation within a state as well as between states by sharing best practices and information and conducting joint projects. Second, adopt a centralized distributed model as the structural model for collaboration in order to benefit from economies of scale in infrastructure investments and

diverse expertise. In this approach, LC could help states to identify and preserve their own records by providing funding, facilitated standards development, and coordination. Third, establish more formal partnership structures between states in order to facilitate communication and secure collaboration and institutional legitimacy against political changes. Fourth, LC could function as a clearinghouse for standards, models, and best practices for digital preservation of state government information in order to facilitate communication and knowledge sharing between states.

As workshop findings suggest, interorganizational and interprofessional collaboration is only one of many important issues involved in the preservation of government digital information. Future research efforts will need to address other problems such as content appraisal and selection, the choice of metadata and preservation standards, sustainable funding, and long-term access to records as well.

5. ACKNOWLEDGEMENTS

This paper is based upon work supported by the U.S. Library of Congress under the National Science Foundation grant # ITR-0205152. Any opinions, findings, conclusions or recommendations expressed in this paper are those of authors and do not necessarily reflect the views of the U.S. Library of Congress or the National Science Foundation.

6. REFERENCES

- [1] Bern bom, G., Lippincott, J., and Eaton, F. Working together: New collaborations among information professionals. *CAUSE/EFFECT*, 22, 2 (1999), 6-9
- [2] Borbinha, J., Kunze, J., Spinazze, A., Mutschke, P., Lieder, H., Mabe, M., Dixson, L., Besser, H., Dean, B., and Cathro, W. Reference models for digital libraries: Actors and roles. *International Journal on Digital Libraries*, 5, 4 (Aug. 2005), 325-331.
- [3] Consultative Committee for Space Data Systems. *CCSDS 650.0-B-1: Reference Model for an Open Archival Information System. Blue Book. Issue 1*. NASA, Washington DC, 2002.
- [4] Dawes, S. S., and Prefontaine, L. Understanding new models of collaboration for delivering government services. *Communications of the ACM*, 46 (Jan. 2003), 40-42.
- [5] Hall, P. Interprofessional teamwork: Professional cultures as barriers. *Journal of Interprofessional Care, Supplement 1* (May 2005), 188-196.
- [6] Hedstrom, M. *It's About Time: Research Challenges in Digital Archiving and Long-term Preservation*. U.S. Library of Congress, Washington DC, Aug. 2003. http://www.digitalpreservation.gov/repor/NSF_LC_Fin al_Report.pdf
- [7] Lippincott, J. K. Working together: Building collaboration between librarians and information technologists (Coalition for Networked Information). *Information Technology and Libraries*, 17, 2. (Jun 1998), 83-87.

- [8] Martin, K., and Reagan, J. *North Carolina State Government Information: Realities and Possibilities*. State Library of North Carolina, Raleigh, NC, Nov. 2003.
<http://statelibrary.dcr.state.nc.us/digidocs/Workgroup/WhitePaper.pdf>
- [9] Matthews, R. J., Burnett, A. E., Cain, C. C., Dow, S. L., McFadden, D. L., and Baish, M. A. *State-by-State Report on Permanent Public Access to Electronic Government Information*. American Association of Law Libraries, Chicago, IL, 2003.
http://www.ll.georgetown.edu/aallwash/State_PPArepo rt.htm
- [10] McGovern, T. J., and Samuels, H. W. Our institutional memory at risk: Collaborators to the rescue. *CAUSE/EFFECT*, 20, 3 (Fall 1997), 19-21, 49-50.
- [11] Pardo, T. A., Cresswell, A. M., Dawes, S. S., Burke, B., Dadayan, L., Embar, S., and Kwon, H., *Building State Government Digital Preservation Partnerships: A Capability Assessment and Planning Toolkit*. Center for Technology in Government, University at Albany, SUNY, Apr. 2005.
http://www.ctg.albany.edu/publications/guides/digital_preservation_partnerships
- [12] Snyder, W. M., Wenger, E., and Briggs, X. de S. Communities of practice in government: Leveraging knowledge for performance. *The Public Manager*, 32, 4 (Winter 2003), 17-22
- [13] U.S. Library of Congress. *Preserving Our Digital Heritage: Plan for the National Digital Information Infrastructure and Preservation Program*. U.S. Library of Congress, Washington DC, Oct. 2002.
<http://www.digitalpreservation.gov/index.php?nav=3&ubnav=1>
- [14] U.S. Library of Congress. *Preservation of State Government digital Information: Issues and Opportunities*. U.S. Library of Congress, Washington DC, Oct. 2005.
http://www.digitalpreservation.gov/repor/states_wkshps .pdf
- [15] Webb, C. *Guidelines for the Preservation of Digital Heritage*. United Nations Educational, Scientific and Cultural Organization, Paris, Mar. 2003.
<http://unesdoc.unesco.org/images/0013/001300/130071 e.pdf>
- [16] Wenger, E. *Communities of Practice: Learning, Meaning, and Identity*, Cambridge University Press, New York, 1998.
- [17] Wenger, E., McDermott, R., and Snyder, W..M. *Cultivating Communities of Practice*, Harvard Business School Press, Boston, 2003
- [18] Wenger, E., and Snyder, W. M. Communities of practice: The organizational frontier. *Harvard Business Review*, 78 (Jan. 2000), 139-146.

Robust Technologies for Automated Ingestion and Long-Term Preservation of Digital Information

Joseph JaJa

Institute for Advanced Computer Studies

Department of Electrical and Computer Engineering
University of Maryland, College Park
301-405-1925

joseph@umiacs.umd.edu

ABSTRACT

In this summary, we present an overview of our DIGARCH project and report on a number of significant advances achieved thus far. In particular, we highlight our contributions to the development of a novel architecture for the Global Digital Format Registry, the design of a highly reliable and scalable deep archive, and the development of the underpinnings of a policy-driven management of preservation processes. Challenges and future plans are briefly outlined.

Categories and Subject Descriptors

H.3.7 [Information Systems] digital libraries

E.2 [Data Storage Representations] object representation

Keywords

Digital preservation, digital archiving, format registry, automated ingestion, management of preservation processes

1. PROJECT OBJECTIVES AND APPROACH

A large portion of the scientific, business, cultural, and government digital information being created today needs to be maintained and preserved for future use of periods ranging from a few years to decades and sometimes centuries. The main goal of this project is to develop technologies for automated ingestion and management of preservation processes for long term preservation of digital information. These technologies will be tested and evaluated on scientific, historical, and educational collections covering widely different technical requirements. The collections include: (i) an archive of videotaped oral histories provided by the Survivors of the Shoah Visual History Foundation; (ii) children's books in their original languages available through the International Children's Digital Library (ICDL: www.icdl.org); (iii) a rich historical collection of photographs, drawings, maps,

charts and documents available through the National Archives' Electronic Access Project; and (iv) a variety of unique earth science data available through the Global Land Cover Facility (www.glc.org).

The foundation of our approach is based on the following principles. The first is to encapsulate properties of content, structure, context, presentation, and preservation within a digital object architecture. The second principle is to separate the management of the digital objects into three levels of abstraction, resulting in a well-defined three-layered architecture. The data layer is responsible for managing the bits across storage systems while the second layer deals with the semantics (metadata) of the data rather than storage and bits. The third layer enables information discovery, search, access, and presentation of the requested digital objects. The data layer also includes a separate support for a deep archive to serve as the ultimate recourse for lost data. The third principle advocates that preservation should be organized as a collaborative endeavor to leverage infrastructure support, share resources and knowledge, and develop community-based efforts.

2. RESEARCH CONTRIBUTIONS

We group our recent contributions under four areas.

2.1 Design of a Digital Global Format Registry

One of the most challenging problems confronting the long term preservation of digital objects is how to handle format obsolescence. Methodologies advocated in the literature include migration, emulation, and standardization to a few common formats. An orthogonal approach, which can exploit progress through any of these methodologies, is to establish a global format registry of digital representation formats[1]. We developed a novel architecture for a Global Digital Format Registry (GDFR) based on scalable, extensible, and secure web technologies. A prototype, called FOCUS (FOrmat Curation Service) [8], has been built and demonstrated to a number of research groups. The new architecture is shown to be more scalable, robust, and comprehensive than any of the previous designs (such as [2],[3],[4]). We are in the process of writing a detailed report that

describes the architecture, testing, and performance of FOCUS. Several groups (including Steve Abrams' group at Harvard and Mackenzie Smith at MIT) have already expressed interest in collaborating with us to further develop the GDFR. For more details about FOCUS, including a demonstration, the reader can check:

<http://www.umiacs.umd.edu/research/adapt/focus/>

2.2 Deep Archive

We have developed a detailed design for a distributed deep archive based on peer to peer technologies using erasure-resilient codes. This new design achieves an extremely high reliability while requiring only a modest amount of additional storage. Briefly, each object is assigned a unique global identifier and is encoded into a set of fragments that are distributed among a number of peers. The encoding is designed so that almost any half of these fragments can be used to fully reconstruct the object. It can be shown that such a scheme guarantees with high probability the integrity and persistence of the distributed deep archive in the presence of system failures and security attacks. We have also developed a placement scheme in such a way to enable a quick retrieval of the appropriate fragments of an object given its ID.

2.3 Automated Ingestion Tools

We have continued our development of automated ingestion tools, which include authoring tools to build the Submission Information Packet (SIP) as per the OAIS model. These tools, grouped under the name PAWN (Producer-Archive Workflow Network) [5,6], enable automated, secure, and scalable distributed ingestion of digital objects into an archive. PAWN was customized for the ingestion of the ICDL collection, and has been used successfully by the ICDL group to demonstrate efficient ingestion. PAWN is currently being more fully tested by the National Archives, and is being used by the Pilot Persistent Archive project, led by the San Diego Supercomputer Center in collaboration with the University of Maryland, NARA, and Georgia Tech.. For more detailed information about PAWN, check:

<http://narawiki.umiacs.umd.edu/twiki/bin/view>

2.4 Policy Driven Management of Preservation Processes

We have started the development of a policy-driven framework based on OWL to manage the preservation processes of an archive. Our initial prototype incorporates policies for replication, refreshing and migration. We expect to complete the development of an architecture and a prototype within the next few months.

3. CHALLENGES AND FUTURE PLANS

The long-term preservation of digital information has to deal with multiple sets of challenges that include social, business, legal, and institutional issues in addition to the development of the appropriate technologies to deal with technology evolution and authenticity preservation. Given that most of the non-technical issues are still open-ended, we have to work with continuously changing sets of assumptions and constraints as well as with evolving technology standards and protocols. Under these circumstances, we have adopted a flexible, robust, and scalable framework based on web technologies and open standards, and we have been working closely with librarians and archivists on significant rich collections. The major challenge remains to show that this approach is indeed flexible enough so as to adapt to the changing requirements in a cost-effective manner. One of our future goals is to develop evaluation techniques to make such an assessment. At the same, we will continue the development of the core areas mentioned in the previous section, with a special emphasis on the architecture and prototype development for policy driven management of preservation processes.

4. REFERENCES

- [1] Abrams, S.L., and Seaman, D. *Towards a Global Format Registry*, IFLA 2003.
<http://www.ifla.org/IV/ifla69/papers/128eAbramsSeaman.pdf>
- [2] JHOVE, JSTOR/Harvard Object Validation Environment
<http://hul.harvard.edu/jhove/>
- [3] Global Digital Format Registry: FRED,
<http://tome.library.upenn.edu/fred/>
- [4] PRONOM, UK National Archives
<http://www.records.pro.gov.uk/pronom/>
- [5] Smorul, M., and JaJa, J. *PAWN: A Novel Ingestion Workflow Technology for Digital Preservation*. Invited talk, ERPLANET Workshop on Workflow, Oct. 13-15, 2004, Budapest, Hungary.
- [6] JaJa, J., Smorul, M., McCall, F., and Wang, Y. Scalable, Reliable Marshalling and Organization of Distributed Large Scale Data onto Enterprise Storage Environments, *Proceedings of the NASA/IEEE Conference on Mass Storage Systems and Technologies*, Monterey, CA, April 2005.
- [7] Moore, R., JaJa, J., and Chadduck, R. Mitigating Risk of Data Loss in Preservation Environments, *Proceedings of the NASA/IEEE Conference on Mass Storage Systems and Technologies*, Monterey, CA, April 2005.
- [8] Geremew, M., Song, S., and JaJa, J. *Using Scalable and Secure Web Technologies to Design a Global Digital Format Registry Prototype: Architecture, Implementation, and Testing*, to appear in *Proceedings of Archiving 2006 Conference*, May 23-26, 2006, Ottawa, Canada.

Building a Demonstration Prototype for the Preservation of Large-Scale Multimedia Collections^{*}

Arcot Rajasekar (PI)
Richard Marciano
Reagan Moore
Chien-Yi Hou
Francine Berman (co-PI)
San Diego Supercomputer Center, Univ. of California, San Diego

Lynn Burstan (co-PI)
Steve Anderson
Mellisa McEwen
Bee Bornheimer
UCSD-TV, Univ. of California, San Diego

Harry Kreisler
UCTV-Berkeley

Brian Schottlaender (co-PI)
Luc DeClerck
Brad Westbrook
Arwen Hutt
Ardys Kozbial
Chris Frymann
Vivian Chu
UCSD Libraries, Univ. of California, San Diego

Abstract

The NSF-DIGARCH is building digital preservation lifecycle management infrastructure for the preservation of large-scale multimedia collections. The infrastructure consists of interfaces to TV production lifecycle systems, metadata definition and capture systems, and a persistent archive workflow which preserves the material in a SRB data grid. Kepler is used to build the workflow.

1. Introduction

The preservation framework development is viewed as a three part process which needs to interact constantly. The first part of this framework is the pre-existing video production lifecycle that should be preserved as much as possible; the second part is the metadata flow, capture and modeling framework that needs to be addressed in order to access, capture and finally preserve the additional material that is needed to complete the preservation packages, and; the third part of the framework is the persistent archive infrastructure and associated workflows [1].

2. Video Production

For nearly 25 years now, Harry Kreisler has been conducting interviews as part of the “Conversations with History” series [2]. Over 230 guests have been

interviewed, including diplomats, statesmen, soldiers, economists, political analysts, scientists, historians, writers, foreign correspondents, activists, and artists. These interviews are one-hour video-taped conversations.

This significant “at risk” collection includes video, audio, text transcripts, web-based material, databases of administrative and descriptive metadata and contains diverse types of data, created at multiple stages within the content production workflow.

Initial “archiving” of the video content has 230 programs in 3 formats:

- digital master files in DV format (.mov files) of typical size 12GB (compressed)
- UCTV broadcasting file in MPEG format of typical size 2GB
- Web archive files in Real Player format of typical size 200 MB

This makes the video content roughly 15GB per show or 230 x 15GB = 3.5TB for the complete collection. When preserving this content, we will replicate the collection in at least two locations, making this a 7TB persistently archived collection

3. Metadata and Modeling

A series of modeling exercises of the existing production workflow were carried out. Two production workflows are described. The first is for creation and transfer of the video taped interviews of the CwH (Conversations with History) program at Berkeley, and its subsequent transfer to the UC/SD TV broadcast studio, which eventually broadcasts and webcasts the program. The second workflow is to model the flow of the audio transcript of the interview produced by CwH

* This research was sponsored the National Science Foundation and the Library of Congress. Views and conclusions contained in this report are the authors' and should not be interpreted as representing the official opinion or policies, either expressed or implied, of the Government, or any person or agency connected with them.

staff. Both workflows reflect the descriptive, technical, and rights metadata that needs to be created during the lifecycle of the interview to successfully manage the interview files for the long term.

3.1. First Work Flow: Capturing / Preserving Video of Interview

The interview is taped by UCB staff. The interview is then described, and the description is forwarded to UCSD-TV with a digital version of the original taping.

UCTV staff enters the description into their FileMakerPro database and augments it wherever useful. UCSD-TV also makes three versions of the original digital file. The files are an edited DV version, a MPG version, and a RealPlayer version.

UCSD-TV outputs an XML record containing preservation descriptive metadata for the interview and preservation technical metadata for each of the content files. This forms the first AIP for preservation.

3.2. Second Work Flow: Capturing / Preserving Transcript of the Video

The interview is transcribed from the video by CwH staff. CwH staff submits an rtf version of the transcript file, along with technical metadata for the file, to SDSC for inclusion in the AIP for the interview.

SDSC verifies the submission and, if all is valid, adds the transcript file and its technical metadata to the AIP for the interview.

These workflows are the basis for the following SIP and AIP models (OAIS-based Submission Information Package and Archival Information Package).

4. Preservation

The preservation of CwH content is based on using data grid technology to manage distributed data. The Storage Resource Broker (SRB) is used as the preservation repository. A central metadata catalog (MCAT) manages preservation metadata for each video file. A dedicated MCAT instance called UCTVStudioArchive was set up. Two additional logical storage resources were registered to store digital video replicas on SAMQfs (uctv-fs1) and HPSS (hpss-sdsc). Also, SRB client software and Kepler scientific workflow software were installed on the eMac machine at UCSD-TV. Finally, a grid brick (srbrick7) with 300GB of disk was configured for the DIGARCH project.

A grid brick [6] is a low-cost commodity disk system to store electronic records. Copies of the electronic records can still be kept on a tape archive at SDSC for

minimizing risk of data loss. Grid Bricks are modular systems that are managed by data grid technology. As additional storage space is needed, additional grid bricks can be added to the data grid, and the electronic records can be automatically distributed across the new storage modules.

5. Acknowledgements

We are working closely with Efrat Jaeger and Ilkay Altintas from the Kepler group and Lucas Gilbert from the SRB group (Jargon) and wish to thank them for their support.

6. Summary

The DigArch project integrates preservation processes on top of a SRB data grid for long-term preservation with a Kepler-based workflow system. The unique aspect of the project is the integration of the preservation processes into a production workflow.

7. References

1. **ICADL 2005, The 8th International Conference on Asian Digital Libraries, December 12-15, 2005, Bangkok, Thailand**, “*Digital Preservation Lifecycle Management for Multimedia Collections*”, Arcot Rajasekar, Reagan Moore, Fran Berman, Brian Schottlaender, <http://www.icadl2005.ait.ac.th/program.htm>
2. Kreisler, H. “Conversations With History”, UC Berkeley, Institute of International Studies, <http://globetrotter.berkeley.edu/conversations/>
3. UCSD-TV, <http://www.ucsd.tv/>
4. Kepler: A System for Scientific Workflows, <http://kepler-project.org/>
5. SRB, Storage Resource Broker, Version 3.1, <http://www.sdsc.edu/dice/srb>, 2004.
6. **Data Grids, Collections and Grid Bricks**, Arcot Rajasekar, Michael Wan, Reagan Moore, George Kremeneck, and Tom Guptill, *20th IEEE/11th NASA Goddard Conference on Mass Storage Systems & Technologies (MSST2003)* San Diego, California, April 7-10, 2003.
7. Developing Data Grid Workflows using Storage Resource Broker and Kepler, Tim Wong, UC-Davis. http://www.thwong.com/documents/SRB_Paper.doc

DIGARCH Project Highlights

Multi-Institutional Testbed for Scalable Digital Archiving

Stephen P. Miller

Scripps Institution of Oceanography
UCSD, 9500 Gilman Drive
La Jolla, CA 02093-0505
(858) 534-1898

spmiller@ucsd.edu

Robert S. Detrick

Woods Hole Oceanographic Institution
360 Woods Hole Road
Woods Hole, MA 02543
(508) 289-3335

rdetrick@whoi.edu

John Helly

San Diego Supercomputer Center
UCSD, 9500 Gilman Drive
La Jolla, CA 92093
(858) 534-5060

hellyj@ucsd.edu

ABSTRACT

This project addresses two major issues of the NSF/Library of Congress DIGARCH program at the same time: digital preservation and inter-institutional collaboration. There is no shortage of at-risk digital media in the physical archives of the Woods Hole Oceanographic Institution (WHOI). Many of the objects have extraordinary value, for both research and education. The cost to recover lost objects is high, at a rate of approximately \$50K per day at sea.

The Scripps Institution of Oceanography (SIO) and the Woods Hole Oceanographic Institution (WHOI) have joined forces with the San Diego Supercomputer Center (SDSC) to build a testbed for multi-institutional archiving of shipboard and deep submergence vehicle data. In addition to the more than 92,000 objects stored in the SIOExplorer Digital Library, the testbed will provide access to data, photographs, video images and documents from WHOI ships, Alvin submersible and Jason ROV dives, and deep-towed vehicle surveys. An interactive digital library interface allows combinations of distributed collections to be browsed, metadata inspected, and objects displayed or selected for download.

1. INTRODUCTION

The SIO/WHOI/SDSC Digital Archive (DIGARCH) project is an effort to construct a testbed for multi-institutional archiving of shipboard and deep submergence vehicle data. This effort aims to extend the successful SIOExplorer project to encompass data from WHOI ships and undersea vehicles. Since this project began in June 2005, a team of scientists, engineers, and librarians has collaborated to develop an implementation plan and strategy. A key goal of the project is to establish methods for effective interoperability of information resources across institutions with complementary scientific objectives but different institutional conventions for managing digital information. The physical resources required have been identified, acquired, and put into service. The base level software from SDSC and SIO has been ported to WHOI, along with initial ingest of a few WHOI-derived data objects. The interfaces needed between current WHOI access tools and the SIOExplorer framework have begun to take shape. Reports and presentations have been made to other agencies and professional groups. During the remainder of this project, efforts will continue to integrate additional WHOI data objects into the framework. A system of Controlled Vocabularies is being developed to allow sensor systems to be added to the

metadata in a scalable fashion, and to insure reliable search results. A preliminary WHOI Web-based front end will be created based on the SIO model. In addition, a second front-end will be implemented to access the same data using WHOI's GeoBrowser model. Efforts will be made to develop an Implementation Document, both for internal quality control, and external use as a guide for other projects. The principal challenge during this period will continue to be overcoming the diversity inherent in the multi-institutional and multi-discipline nature of modern ocean sciences research.

2. ACCOMPLISHMENTS

Shortly after this project began in June 2005, the first major accomplishment was to convene a three-day meeting of the PI's and other personnel from SIO, WHOI, and SDSC. During this meeting, discussions were geared toward reaching a common census of tasks to be done, and to establishing an understanding of what each participant could contribute. The end product of this meeting consisted of a detailed time and task list, which has been subsequently used to maintain forward progress. To further coordinate efforts, a Plone-based, collaborative web site was initiated, linking personnel, documentation, and planning.

One major accomplishment that quickly followed was the acquisition of the appropriate computer hardware (server and RAID-based storage) needed to maintain commonality between WHOI and SIO developers. Nearly identical systems have been installed at both institutions, and procedures and conventions established to ensure interoperability.

The migration of the SIOExplorer software framework to WHOI began shortly after the computing resources were put on-line. Much of underlying architecture utilizes mature, well-supported, open software systems (PostgreSQL, SDSC Storage Resource Broker, and Apache web server), which moved effortlessly. Those sub-systems specific to SIOExplorer (metadata cataloging and metadata template file and related controlled vocabulary definitions, manipulation of Arbitrary Digital Objects, procedures for data and metadata harvesting) were more carefully migrated, as these needed to be adapted to match the needs of WHOI generated data products [1-6].

As the WHOI-based system began to take shape, work commenced on adapting legacy WHOI tools to interoperate with the SIOExplorer framework. Two tools in particular, the Alvin Framergrabber and the Jason Virtual Control Van, play a critical role in allowing scientists to access still and video imagery from

WHOI deep submergence vehicles. Both tools make use of the WHOI GeoBrowser [7] technology, so efforts have centered on this important link. In addition, a separate effort has begun to create the link between WHOI GIS-derived, web-based mapping systems and the SIOExplorer framework.

3. RESEARCH CONTRIBUTIONS

A number of presentations have been made during this project, in order to both disseminate information concerning the DIGARCH effort itself, and to proselytize the methodology and concepts of multi-institutional cooperation to peer marine science institutions. The presentations can be downloaded from <http://gdc.ucsd.edu:8080/digarch/about-project/presentations/>.

Members of the team participated in DIGARCH program discussions in the dg.o2005 meeting in Atlanta in May 2005, a progress report at the Library of Congress in November 2005, and at the NDIPP digital partners meeting in Berkeley in January 2006. Two presentations were made at the Fall 2005 Meeting of the American Geophysical Union. The first [8] highlighted the overall goals of the DIGARCH effort, principally as it relates to the broad science community. The second [9] presentation showed more of the technical aspects of the DIGARCH project.

As part of the continuing development of our digital library techniques for shipboard data, the system was field-tested on the R/V LAURENCE M. GOULD by Helly, in a study of Antarctic icebergs conducted in December 2005. The results were good and this represents the fifth oceanographic research ship on which this methodology has been deployed.

4. THE COMING YEAR

During the remainder of the project, the primary focus will be to extend the current effort in defining the procedures needed to create metadata and Arbitrary Digital Objects to a wider selection of WHOI shipboard and vehicle data. The pace of this effort should accelerate as a broader range of personnel acquires more expertise. In addition, joint efforts between SIO and WHOI will result in the adoption of a controlled vocabulary needed to implement advanced searching techniques.

Work will also be undertaken to finalize the linking of the SIOExplorer framework with current WHOI access tools. Since a high proportion of the value of WHOI deep submergence data lies in the extensive video imagery domain, this link will be critical for acceptance to the science community.

It is also expected that during the remaining months, efforts will be made to begin defining the requirements and specifications of a WHOI Web Site geared toward access of WHOI data. Efforts will also be needed to begin writing and assembling of Implementation Documentation, to serve both as an internal quality control mechanism, and as an impetus for use beyond the current two institutions.

5. ACKNOWLEDGEMENTS

These efforts have been supported by the Digital Archiving and Preservation Program (DIGARCH) of NSF and the Library of Congress under award NSF IIS 0455998.

6. REFERENCES

- [1] Helly, J., in *Hydroinformatics: Data Integrative Approaches in Computation, Analysis, and Modeling* M. Praveen Kumar, Momcilo Markus, Jay C Alameda, Peter Bajcsy, Ed. (Taylor and Francis, London, England, 2005) pp. 552.
- [2] Helly, J., H. Staudigel, A. Koppers (2003). Scalable Models of Data Sharing in the Earth Sciences, *Geochemistry, Geophysics, Geoscience*, 1010, doi:10.1029/2002GC000318 4(1): 14.
- [3] Staudigel, H., Helly, J., et al. (2003), Electronic data publication in geochemistry, *Geochemistry, Geophysics, Geoscience* DOI number 10.1029/2002GC000314.
- [4] Helly, J., New concepts of publication, *Nature*, 393 (1998), pp. 107.
- [5] Helly, J., T. T. Elvins, D. Sutton and D. Martinez, (1999), A Method for Interoperable Digital Libraries and Data Repositories, *Future Generation Computer Systems*, Elsevier, 16 (1999), pp. 21-28.
- [6] Helly, J., T. T. Elvins, D. Sutton, D. Martinez, S. Miller, S. Pickett and A. M. Ellison, (2002), Controlled Publication of Digital Scientific Data, *Communications of the ACM*, 45(5):May 2002, pp. 97-101
- [7] Lerner, S., Maffei, A., (2001), “4DGeoBrowser: A Web-Based Data Browser and Server for Accessing and Analyzing Multi-Disciplinary Data”, Woods Hole Oceanographic Institution, Technical Report, WHOI-2001-13, October 2001.
- [8] Detrick, R.S., D. Clark, A. Gaylord, R. Goldsmith, J. Helly, P. Lemmond, S. Lerner, A. Maffei, S. P. Miller, C. Norton, B. Walden, (2005), IN44A-07, WHOI and SIO (I): Next Steps toward Multi-Institution Archiving of Shipboard and Deep Submergence Vehicle Data, *EOS, Trans. Amer. Geophys. Union*,
- [9] Helly, J., A. Maffei, P. D. Clark, R. Detrick, A. Gaylord, R. Goldsmith, P. Lemmond, S. Lerner, S. Miller, C. Norton, B. Walden, (2005), IN51A-0306, WHOI and SIO (II): Next Steps toward Multi-Institution Archiving of Shipboard and Deep Submergence Vehicle Data, *EOS, Trans. Amer. Geophys. Union*.

SESSION 7C

SPATIO TEMPORAL AND GIS

Moderator

Andrew Philpot, University of Southern California, USA

Titles and Authors

Voting Prediction Using New Spatiotemporal Interpolation Methods
Gao, Jun; Revesz, Peter

Scalable Data Collection and Retrieval Infrastructure for Digital Government Applications
Samet, Hanan; Golubchik, Leana

Automatic Alignment of Vector Data and Orthoimagery for The National Map
Knoblock, Craig A.; Shahabi, Cyrus; Chen, Ching-Chien; Usery, E. Lynn

National Large-Scale Urban True Orthophoto Mapping and Its Standard Initiative
Zhou, Guoqing; Xie, Wenhan; Benjamin, Susan; Fegeas, Robin G.; Simmers, John; Cluff, Hap;
Lei, Y.; Foust, Jeanne

Voting Prediction Using New Spatiotemporal Interpolation Methods*

Jun Gao and Peter Revesz

Department of Computer Science and Engineering
University of Nebraska-Lincoln

Lincoln, NE 68588, USA

jgao,revesz@cse.unl.edu

ABSTRACT

Most spatial and spatiotemporal interpolation methods give back a surface function as the result. Instead of that we consider interpolation methods that yield a single value as the final result. *Voting prediction* is a natural example that requires this type of spatiotemporal interpolation, because the final result is the total percentage vote for a party or candidate. We propose a new spatiotemporal interpolation method for voting prediction and similar problems. The approach can also be used in election data verification for effective government. We test the new method using USA presidential election data from the states of California, Florida, and Ohio between 1972 and 2004. The experimental results show that our method can produce comparatively precise predictions (e.g., the difference between prediction and actual result is 1.09% for Florida in 2004).

Categories and Subject Descriptors

G.1.1 [Numerical Analysis]: Interpolation—*Interpolation models, interpolation formulas*; J.1 [Administrative Data Processing]: Government

General Terms

Algorithms, Experimentation, Measurement

Keywords

Voting prediction, spatiotemporal, interpolation

1. INTRODUCTION

Spatial and spatiotemporal interpolations are important in many problems, such as, geographically distributed statistics for agricultural productions, disease prevalence, pollution levels, soil types, precipitation, and temperatures. Spatiotemporal interpolation is used to interpolate the original point-based Standardized Precipitation Index (SPI) data in

*This research was supported in part by NSF grant EIA-0091530 and a NASA Space and EPSCoR grant.

a drought online analysis system [23]. These usually require the estimation of the unknown values at unsampled location-time pairs and yield as the final result a surface function. In contrast, we consider spatiotemporal interpolation methods that require only a single value as the final result. Aiming at effective government we choose predicting presidential election as the main application in this study.

Most presidential election forecasting models use *multi-variate ordinary least squares regression*, a common statistical method in the social sciences [10]. Those models compare calculations from previous elections of such independent variables as presidential popularity and economic growth with their current values to estimate the result in a future election. Among the simplest forecasting models are several that predict national two-party vote shares using time series data and sets of explanatory variables. Campbell and Wink use just two predictor variables, a trial-heat poll and second quarter GDP growth in the year of the election [2]. Lewis-Beck and Rice use a similar specification, but add variables capturing recent partisan trends [15].

As pointed out by Chappell [4], although the national models is widely accepted, predicting shares of the popular vote should not be the principal objective when the election winner is selected according to outcomes in the individual states. Several models such as the models proposed by Rosenstone[19] and Campbell [3] are designed for forecasts at the state-level. These models examine election outcomes across both states and time, using a mixture of national-level and state-level variables as explanatory variables.

Although both the national-level and state-level models introduced above are frequently cited for their use in forecasting and the accuracy is admirable, most of them share limitations. For example, the choice of factors to include in the model adds to the uncertainty. The decision to include one set of variables, such as presidential popularity and growth in GNP, rather than another, such as the rate of inflation and unemployment, changes the prediction outcome [10]. Also most models are limited by the lack of historical information on the relationship between political and economic fundamentals and elections [10]. In our research we turn the direction into the historical election data itself as the basis of spatiotemporal interpolations without a set of variables.

A key issue is the choice of an appropriate interpolation method for a given input data [1, 17]. Inverse distance

weighting (IDW) [13, 18, 20, 21], kriging [7], shape functions [16], splines [9], and trend surface analysis [24] are some of the common spatial interpolation methods. We propose a novel and comparatively simple spatiotemporal interpolation method as a combination of spatial and temporal interpolation methods to predict the election at the state-level.

The rest of the paper is organized as follows. Section 2 reviews inverse distance weighting, which is a popular spatial interpolation method. We also use the IDW method in this study. Section 3 describes our spatiotemporal interpolation method. Section 4 describes the experimental methods and results. Finally, Section 5 presents some ideas for future work.

2. INVERSE DISTANCE WEIGHTING

Distance-based weighting methods have been used to interpolate spatial data by many authors, for example, by Legates and Willmont [13] and Stallings et al. [21]. The main assumption of IDW is that if A , B and C are three different locations, such that A is closer to B than to C , then the value we are interested in (temperature, precipitation, percentage of voters preferring a particular candidate, etc.) is also closer between A and B than between A and C . Hence, if the value at location A is unknown, while the values at locations B and C are known, then the value at B should be more important than the value at C in estimating the value at A .

The relative importance of the known values is reflected by the weights assigned by the IDW method to them. In the IDW method the sum of the weights is equal to 1, and the weights are assigned proportionally to the *inverse of the distance* between the known and unknown locations.

Let λ_i be the weight for the individual location, and y_i the variable observed in the sampled location.

IDW interpolations are of the form [11]:

$$y = \sum_{i=1}^N \lambda_i \cdot y_i \quad (1)$$

$$\lambda_i = \frac{\left(\frac{1}{d_i}\right)^p}{\sum_{k=1}^N \left(\frac{1}{d_k}\right)^p} \quad (2)$$

For simplicity in the following we assume that $p = 1$. Therefore,

$$\lambda_i = \frac{\frac{1}{d_i}}{\sum_{k=1}^N \frac{1}{d_k}} \quad (3)$$

EXAMPLE 1. Assume that $A = (5, 0)$, $B = (0, 0)$, and $C = (20, 0)$ and the value at A is unknown but the values at B and C are 100 and 200, respectively. Then, the number

of known points is $N = 2$. We use the subscripts B and C instead of numbers in this simple example. We can calculate that:

$$\lambda_B = \frac{\frac{1}{5}}{\frac{1}{5} + \frac{1}{15}} = 0.75 \quad \lambda_C = \frac{\frac{1}{15}}{\frac{1}{5} + \frac{1}{15}} = 0.25$$

Hence the value of A will be interpolated based on B and C to be:

$$y_A = \lambda_B y_B + \lambda_C y_C = 0.75 \times 100 + 0.25 \times 200 = 125$$

Note that since point C is three times more distant than B is from point A , the weight λ_C is only a third of the weight λ_B . Hence y_A is much closer to y_B than to y_C .

3. NEW SPATIOTEMPORAL INTERPOLATION METHODS

Now we describe a new spatiotemporal interpolation method which is a combination of a spatial interpolation method with a temporal interpolation method. For the spatial interpolation part, we consider to use the IDW method as described in Section 2. We choose IDW because its ease of use and low computation charge [5]. For the temporal interpolation part, we consider two methods as described in Section 3.1.

For any location C , let $E_{t,C}$ be the estimated value using any chosen temporal interpolation method, and $E_{s,C}$ the estimated value using any spatial interpolation method, α_C the weight of $E_{t,C}$, and β_C the weight of $E_{s,C}$. We calculate the overall estimation value E_C for location C as follows:

$$E_C = \alpha_C \times E_{t,C} + \beta_C \times E_{s,C} \quad (4)$$

where $\alpha_C + \beta_C = 1$ and $0 \leq \alpha_C, \beta_C \leq 1$.

In interpolating the percentage vote for a given party in some county C for which we do not have information, we would naturally like to rely on the percentage votes in its neighboring counties if those values are known. Here we should notice that since election voting is not like some GIS applications like minimum or maximum temperatures in a weather station, where some data are missing because of broken instruments or data processing mistakes, hence interpolation is needed to find the replacing values. For election voting, it is very unlikely that previous voting result can not be found. And what people are most interested in is who will win in the coming election. Therefore, instead of doing a interpolation, we use our method to do a prediction.

Now we discuss how to determine $E_{t,C}$, $E_{s,C}$, α_C , and β_C in the following.

3.1 Temporal methods to determine $E_{t,C}$

3.1.1 Inverse linear temporal method

This is a variant of the IDW methods that measures “distance” in terms of time difference instead of spatial difference. That is, it treats time as a third dimension. Following

the IDW method, the weights are assigned proportional to the inverse of the time difference, and again we assume that $p = 1$.

3.1.2 Inverse exponential temporal method

After some experimentation we realized that time is special and the inverse linear temporal method does not yield good results. Increasing p to a small constant 2 or 3 also does not yield a good result. Hence, we introduce another method that assigns weights that decrease exponentially with the time difference, i.e., if we look back in time n years and have one data in each of the past n years, then the weight of the data i years back in time will be $\frac{1}{2^i}$ for $1 \leq i \leq (n-1)$ and $\frac{1}{2^{n-1}}$ for n years back. Note that the last two weights will be the same and with this rule the sum of the weights is still 1.

Consider predicting the outcome of the USA presidential election of 2004 based on six previous election results, namely the presidential election votes in 2000, 1996, 1992, 1988, 1984, and 1980.

For inverse linear temporal interpolation, we use the time distance of one as the distance between two continuous USA presidential elections (even though it means four years). Hence we get the weights:

$$\lambda_i = \frac{\frac{1}{i}}{\frac{1}{1} + \frac{1}{2} + \dots + \frac{1}{6}} = \frac{1}{2.45 \times i}$$

Using inverse exponential temporal interpolation, we assign the weights to the outcome of these elections (at any city or voting district) as $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}, \frac{1}{32}$, respectively. The weight $\frac{1}{32}$ occurring twice is to keep the sum of the weights still 1.

3.2 Spatial methods to determine $E_{s,C}$

As stated before, IDW is a very popular interpolation method, we use it as the spatial part for our method. However, a problem arises. For example, suppose we are back in November 2004 and want to predict the percentage vote for John Kerry in the 2004 USA presidential election in Alachua county, Florida. It is not reasonable to use the actual votes in Bradford, Clay, Columbia, Gilchrist, Levy, Marion, Putnam, and Union, which are the neighboring counties of Alachua, because those votes are not known yet. A possible solution is to use the estimated data in the neighboring counties, which can be created by many methods such as our inverse linear or inverse exponential temporal methods.

When we use the IDW method, we consider two versions. One is the IDW method with uniform distances and the other with real distances.

3.2.1 IDW with uniform distances

Suppose we want to predict the votes for county C , which has the following neighboring counties, N_1, N_2, \dots, N_k . We assume all the distances between counties C and N_i , $1 \leq i \leq k$, are the same. Hence by Equation (3) each neighbor N_i has exactly the same weight $\lambda_i = \frac{1}{k}$, $1 \leq i \leq k$.

Table 1: Latitude and longitude of centroid of 67 counties of Florida, USA

County name	Latitude	Longitude
Alachua	29.676436	-82.379953
Baker	30.287517	-82.236268
Bay	30.219170	-85.638788
...		
Wakulla	30.144620	-84.366174
Walton	30.637995	-86.155962
Washington	30.630591	-85.638396

3.2.2 IDW with real distances

When considering the real distance between counties C and N_i , $1 \leq i \leq k$, we calculate the distances between the centroid of counties C and N_i , $1 \leq i \leq k$. Because of the near-spherical shape of the Earth, calculating an accurate distance between two points requires the use of spherical geometry and trigonometric math functions. In this study, we use the formulae introduced by Weisstein [22],

$$distance = \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2 + (z_1 - z_0)^2} \quad (5)$$

where for $0 \leq i \leq 1$

$$\begin{aligned} x_i &= R \times \cos(long_i) \times \sin(90^\circ - lat_i) \\ y_i &= R \times \sin(long_i) \times \sin(90^\circ - lat_i) \\ z_i &= R \times \cos(90^\circ - lat_i) \\ \text{and } R &= 6368KM \end{aligned}$$

The latitude and longitude of the centroid of a county are shown in Table 1, which is obtained from the official website <http://www.census.gov>.

Using the above distances we can get the weights of neighboring counties by Equation (3). Combined with the data of neighboring counties, we can use Equation (1) to estimate the votes for county C .

An interesting aspect is that in states with long and narrow shapes, such as Florida, there are fewer neighbors on average for each county than in counties with a more round shape such as Ohio. Therefore, we were concerned that the overall shape of a state can influence heavily the accuracy of our spatiotemporal interpolation method. Hence we choose three states, that is, Florida, Ohio, and California, with very different shapes as our test cases.

In each of our three test states, there are counties that have additional neighbors in other states. For example, some counties in Florida are neighbors of some counties in Georgia. However, we did not count neighbors in other states, because we did not have available data for them. Presumably the accuracy of our interpolation methods can be further improved by counting those neighbors too.

Table 2: d_i and σ of 67 counties of Florida, USA

d_i in each county	00/96	96/92	92/88	88/84	84/80	σ
Alachua	1.353543	4.287756	0.781069	2.403800	5.863571	2.937948
Baker	4.927593	5.031435	0.896374	0.045522	24.17032	7.014248
Bay	0.951147	4.895554	1.633311	2.241453	11.67629	4.279550
...						
Wakulla	2.071604	8.078953	1.023610	1.278107	16.490388	5.7885324
Walton	3.626663	5.728941	1.235268	3.937664	20.734451	7.0525974
Washington	3.204958	5.745716	0.326218	3.265001	18.464455	6.2012696

3.3 Determine α_C and β_C

3.3.1 Step function

When we consider this new method, the most natural way to determine α_C and β_C is a step function as shown in Figure 1. In a step function, we find some parameter σ_C and fix some threshold value θ (Details about σ_C and θ are in the following paragraphs). If $\sigma_C < \theta$, then we set $\alpha_C = 1$ and $\beta_C = 0$, which enforces that we use the temporal interpolation method; and if $\sigma_C \geq \theta$, then we set $\alpha_C = 0$ and $\beta_C = 1$, which enforces that we use the IDW spatial interpolation method. In summary,

$$\begin{cases} \alpha_C = 1, \beta_C = 0 & \text{if } \sigma_C < \theta \\ \alpha_C = 0, \beta_C = 1 & \text{if } \sigma_C \geq \theta \end{cases} \quad (6)$$

σ_C and θ are considered according to a specific application. For example, when we apply the method to USA presidential election data, we choose σ_C as the changes in the vote percentages of all pairs of subsequent presidential elections for a county C . We choose θ as a constant, say 1%, 2% and so on. Intuitively, a smaller σ_C means that the values in a county C are more consistent over time, hence we can rely more on the temporal interpolation method, which means that we should increase α_C and decrease β_C .

Let $M_{t,C}$ be the absolute difference between the temporal estimation value and the actual data at location C . Similarly, let $M_{i,C}$ be the absolute difference between the IDW estimation value and the actual data. If most counties with $\sigma_C < \theta$ have $M_{t,C} < M_{i,C}$ while most counties with $\sigma_C \geq \theta$ have $M_{t,C} \geq M_{i,C}$, then the step function makes an ideal choice. Intuition would suggest that $M_{t,C}$ and $M_{i,C}$ are independent, hence if a temporal method is more reliable because σ_C is small, then it is also usually the case that $M_{t,C} < M_{i,C}$.

3.3.2 Linear function

In addition to the step functions, we also experimented with linear functions of the form $\alpha = c\sigma + d$ with different values for the constants c and d . However, the linear functions did not work as well as the step functions. One likely explanation is that the temporal and IDW methods give similar variations for most counties, that is, when the temporal estimation value is higher (or lower) than the original data,

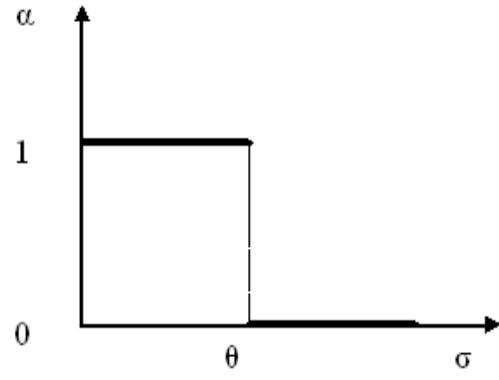


Figure 1: Step function

then the IDW estimation value is also higher (or lower). That makes it difficult to find a good linear function.

3.3.3 σ_C in the election data

Suppose we would like to predict the outcome of the USA presidential election of 2004 in Alachua, Florida. Let us look at how to calculate $\sigma_{Alachua}$.

Let P_{year} be the percentage vote for the democratic candidate in the given year in Alachua and use P_{00} instead of P_{2000} and so on. We have $P_{00} = 55.249682\%$, $P_{96} = 53.896139\%$, $P_{92} = 49.608382\%$, $P_{88} = 48.827313\%$, $P_{84} = 46.423513\%$, and $P_{80} = 52.287084\%$.

Let d be the absolute difference between two continuous USA presidential elections, then $d_1 = |P_{00} - P_{96}|, \dots, d_5 = |P_{84} - P_{80}|$. That is, $d_1 = |55.249682\% - 53.896139\%| = 1.353543\%$, $d_2 = 4.287756\%$, $d_3 = 0.781069\%$, $d_4 = 2.4038\%$, and $d_5 = 5.863571\%$.

Hence we get:

$$\sigma_{Alachua} = \frac{d_1 + d_2 + d_3 + d_4 + d_5}{5} = 2.937948\%$$

Table 2 gives d_i and σ of six counties of the state of Florida.

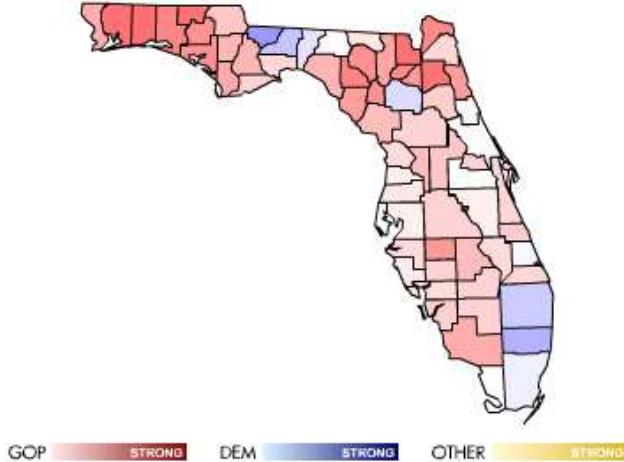


Figure 2: 2004 presidential votes by county in Florida (from <http://www.cnn.com>)

We calculated similarly the σ for the remaining 61 counties in Florida, but we do not show them for space limitations.

4. EXPERIMENTAL METHODS AND RESULTS

4.1 USA presidential election data sets

As stated before, in order to test our idea, we used the USA presidential election data for the states of California, Florida, and Ohio. For Florida, the data is obtained from the official website [25], which is maintained by the Florida Division of Elections and contains a comprehensive USA presidential voting data for 67 different counties in Florida between 1980 and 2004. Table 3 shows a part of the post-calculated data. The map in Figure 2 shows the 2004 presidential votes for each county in Florida. For California and Ohio, the data is obtained from [14], for the time period between 1972 and 2004. We estimated the votes for the 2004 democratic candidate for USA president (John Kerry) in those three states using our new method and compared them with the actual votes.

4.2 Prediction procedures

We tried out both the inverse linear and the inverse exponential temporal methods as described in Section 3.1. We used the two versions of the IDW methods described in Section 3.2 as our spatial interpolation method.

Once we get the temporal and spatial interpolation values, we apply Equation (4) to calculate the final estimation value. We test step functions to find the best estimation parameters α , β , and θ . For the threshold parameter θ we tried the ten values 1%, 2%, 3%, ..., 10%.

4.3 Evaluation method

Several measures are suitable for experimentally comparing the accuracy of interpolation methods. We use mean ab-

Table 3: Votes for 2000 USA presidential election in 67 counties of Florida, USA

County name	Total votes	Votes for Republican candidate	Votes for Democratic candidate
Alachua	85,757	34,135	47,380
Baker	8,155	5,611	2,392
Bay	58,876	38,682	18,873
...			
Wakulla	8,587	4,512	3,838
Walton	18,323	12,186	5,643
Washington	8,026	4,995	2,798

solute error (MAE) and root mean square error (RMSE).

$$MAE = \frac{\sum_{i=1}^N |F_i - A_i|}{N} \quad (7)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (F_i - A_i)^2}{N}} \quad (8)$$

where

F_i : Prediction value.

A_i : Actual measurement.

N : Number of data.

Let $VPstate_e$ be the estimated statewide vote percentage for a given party. Similarly, let $VPstate_a$ be the actual statewide vote percentage for a given party.

$$VPstate_e = \frac{\sum E_i \times V_i}{\sum V_i} \quad (9)$$

where

E_i : Estimated vote percentage for a given party in county i .

V_i : The number of all voters in county i .

Then we can calculate the error of statewide total vote percentage (TE), which is a more interesting measure in the voting prediction area.

$$TE = |VPstate_e - VPstate_a| \quad (10)$$

EXAMPLE 2. Assume that a state S has three counties A , B , and C . For some election the numbers of all voters in counties A , B , and C are 1000, 2000, and 3000, respectively. The estimated vote percentages for a given party in counties

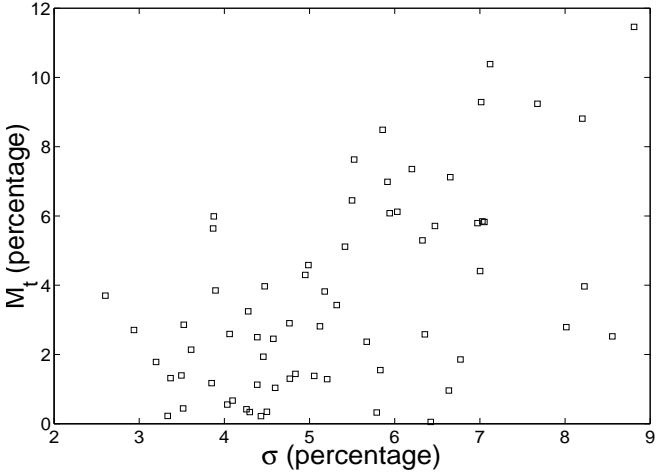


Figure 3: M_t of step function

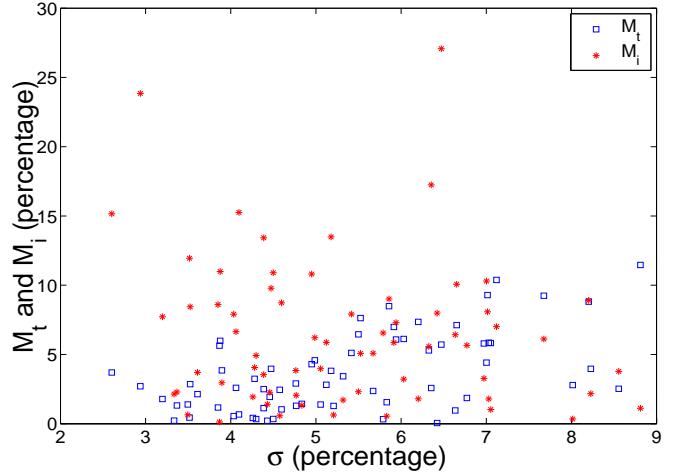


Figure 5: M_t and M_i of step function

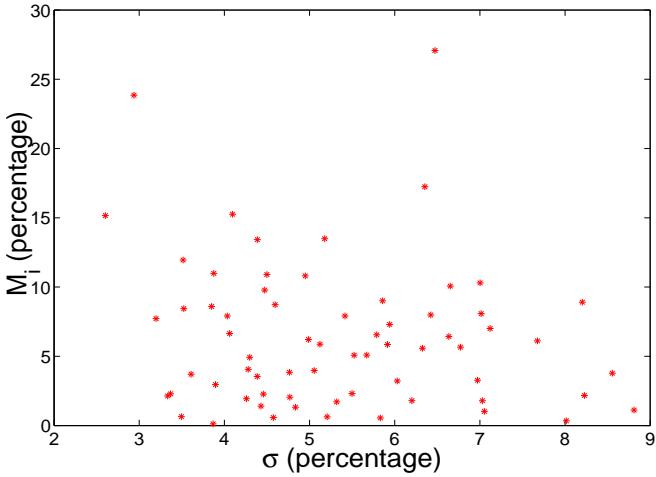


Figure 4: M_i of step function

A , B , and C are 40%, 50%, 60%, respectively. And the actual vote percentage for a given party in state S is 58%. We can calculate that:

$$TE = \left| \frac{40\% \times 1000 + 50\% \times 2000 + 60\% \times 3000}{1000 + 2000 + 3000} - 58\% \right| = 4.7\%$$

4.4 Evaluation

Figures 3-5 give the intuition for step functions, based on 67 counties in Florida. The x-axis is σ in each figure, while the y-axis is M_t in Figure 3, M_i in Figure 4, and their weighted linear combination in Figure 5. It can be seen that for this data set, 7% seems a reasonable threshold since most counties with $\sigma < 7\%$ have $M_t < M_i$, and most counties with $\sigma \geq 7\%$ have $M_t \geq M_i$. The experiments proved true our intuition.

Table 4 records our experimental results. We can see that the performance of spatiotemporal step functions and inverse exponential temporal methods is the best, getting com-

paratively precise predictions, especially in predicting the 2004 USA presidential election in Florida. Spatiotemporal step functions (with $\theta = 7\%$) predict for the 2004 USA presidential election, the democratic candidate (John Kerry) will win 46.00% votes in Florida, and the actual result is 47.09%, hence the discrepancy (TE) is only 1.09%. This contrasts favorably with a CNN poll which predicted only 42% for John Kerry shortly before the election [26], i.e., it had a TE of more than 5%. Let us look at the results of the presidential election forecasting models. For example, when the trial-heat poll model predicts Bush's vote in Florida based on the polls of the week between Oct 25 and Nov 1, TE is 2.2% [6].

The experimental results for California and Ohio are also impressive. Inverse exponential temporal method shows slightly better performance, TE is 3.46 and 3.18 in California and Ohio, respectively. Let us look at the prediction of Bush's vote in exit polls. TE is 1.5 in California and 4.4 in Ohio, respectively [6]. For all three states, MAE and RMSE are reasonably low.

The experiment shows that the difference between the two versions of IDW methods with uniform distances and real distances are extremely small in our case. Therefore, the much more complicated standard IDW method using exact distances can be simplified by IDW method with uniform distances using topological neighbors without any significant change in the accuracy of the result.

Figure 6 shows the estimated vote percentages using step functions and the actual vote percentages over 67 counties in Florida. We can see that for most counties the discrepancy is low and it almost disappears for several counties.

5. CONCLUSION AND FUTURE WORK

The experimental results show that our new spatiotemporal interpolation method can be a basis for an effective voting prediction system. Of course, any real voting prediction system would need to be fine-tuned by considering many additional variables, such as a candidate's expenditures, gender,

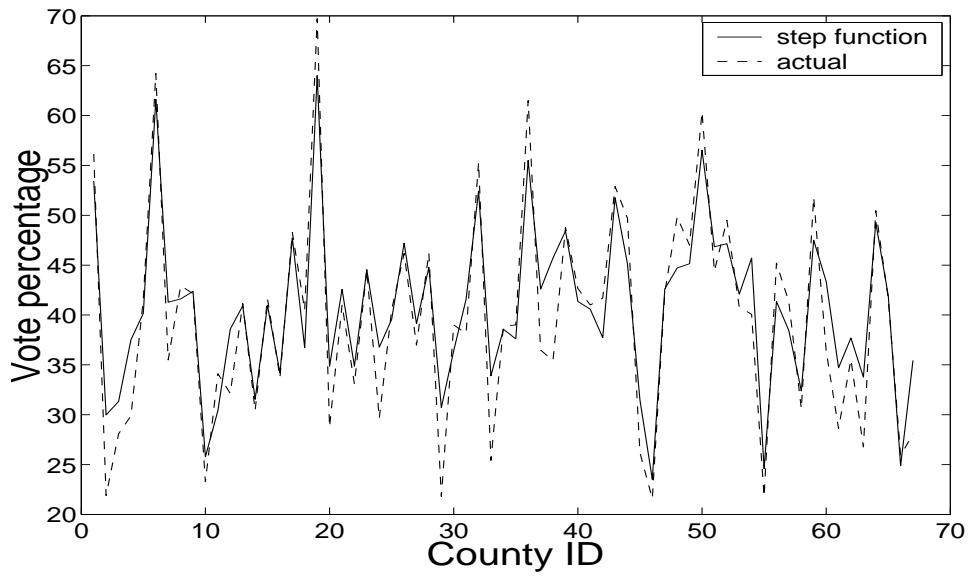


Figure 6: Estimated vote percentages using step functions and actual vote percentages over 67 counties in Florida, USA

Table 4: Comparison of step function, temporal and IDW methods

	California 2004			Florida 2004			Ohio 2004		
	TE	MAE	RMSE	TE	MAE	RMSE	TE	MAE	RMSE
<hr/> IDW using uniform distances <hr/>									
Spatial IDW	8.65	11.60	9.67	4.88	7.98	9.05	8.75	11.31	7.60
Spatiotemporal Step Function ($\theta = 7\%$)	3.49	4.51	6.26	1.09	2.40	5.18	3.57	4.37	3.57
Spatiotemporal Step Function ($\theta = 8\%$)	3.55	4.77	6.38	1.10	2.40	4.72	3.89	4.66	3.88
Spatiotemporal Step Function ($\theta = 9\%$)	3.49	4.51	6.26	1.10	2.39	4.61	3.27	4.05	3.14
<hr/> IDW using real distances <hr/>									
Spatial IDW	8.02	11.33	9.33	3.51	6.62	8.64	8.83	11.27	7.45
Spatiotemporal Step Function ($\theta = 7\%$)	3.58	4.63	6.83	1.10	2.39	4.84	3.45	5.06	4.88
Spatiotemporal Step Function ($\theta = 8\%$)	3.54	4.54	6.32	1.11	2.39	4.69	3.78	4.56	3.71
Spatiotemporal Step Function ($\theta = 9\%$)	3.50	4.51	6.03	1.11	2.39	4.59	3.25	4.03	3.10
<hr/> Temporal Inverse Linear									
Temporal Inverse Exponential	5.46	6.66	7.25	2.68	3.81	5.12	4.10	5.09	3.74
Temporal Inverse Exponential	3.46	4.48	6.01	1.10	2.39	4.59	3.18	3.99	3.10

incumbency, and the interaction affects of those parameters. However, it is extremely interesting and encouraging that by combining a temporal and a spatial interpolation method, which in themselves are not too sophisticated, already yields prediction values that are more accurate than the results —published in various newspapers in the run-up to the elections— of much more sophisticated prediction systems. Hence our vote prediction system has a significant potential that we plan to exploit by factoring in more variables.

As this approach produced both county-level and state-level results, it can be used by election agencies in election data verification for effective government. We can compare the collected election results with the estimates at the county-level and identify possible suspected data when there is significant difference between them.

In the future, we also plan to look at other problems that require a single value as the outcome of the interpolation problem. For example, an aggregate health statistics, such as the number of persons infected with various specific diseases in a state or country would be another natural problem to look at. Another would be to predict human population changes in a country or worldwide. Both of these are known to be hard problems. For example, there are widely different values for the total number of AIDS cases predicted using different models or the predicted total human population in the world. By improving the estimation accuracy of these and similar types of problems, we can help governments and international health and environmental agencies to be better prepared in the future.

6. REFERENCES

- [1] P. A. Burrough and R. A. McDonnell. *Principles of Geographical Information Systems*. Oxford University Press, 1998.
- [2] J. Campbell and K. Wink. Trial-heat forecasts of the presidential vote. *American Politics Quarterly*, 18:251-269, 1990.
- [3] J. Campbell. Polls and votes: the trial heat presidential election forecasting model, certainty, and political campaigns. *American Politics Quarterly*, 24: 408-433, 1996.
- [4] H. W. Chappell. Forecasting Presidential Elections in the United States. Entry Prepared for the Encyclopedia of Public Choice, Charles Prowley and Friedrich Schneider, eds., Springer, 2004.
- [5] F. Collins and P. Bolstad. A Comparison of Spatial Interpolation Techniques in Temperature Estimation. In *Proc. of the Third International Conference/Workshop on Integrating GIS and Environmental Modelling*, 1996.
- [6] A. Cuzán, S. Armstrong, and R. Jones. Combining Methods to Forecast the 2004 Presidential Election: The Pollyvote. *mimeo*, University of Pennsylvania, 2005.
- [7] C. V. Deutsch and A. G. Journel. *GSLIB: Geostatistical Software Library and User's Guide*, 2nd ed., Oxford University Press, 1998.
- [8] R. Fair. Econometrics and presidential elections. *Journal of Economic Perspectives*, 3:89-102, 1996.
- [9] J. E. Goodman and J. O'Rourke, editors. *Handbook of Discrete and Computational Geometry*. CRC Press, New York, 1997.
- [10] J. P. Greene. Forecasting Follies. *The American Prospect*, vol 4 no. 15, 1993.
- [11] K. Johnston, J. M. V. Hoef, K. Krivoruchko, and N. Lucas. *Using ArcGIS Geostatistical Analyst*. ESRI Press, 2001.
- [12] G. Kramer. Short-term fluctuations in U.S. voting behavior. *American Political Science Review*, 65:131-143, 1971.
- [13] D. R. Legates and C. J. Willmont. Mean seasonal and spatial variability in global surface air temperature. *Theoretical Application in Climatology*, 41:11-21, 1990.
- [14] D. Leip. Dave Leip's Atlas of U.S. Presidential Elections. <http://www.uselectionatlas.org>, 2005.
- [15] M. Lewis-Beck and T. Rice. *Forecasting Elections*. Congressional Quarterly Press, 1992.
- [16] L. Li and P. Revesz. Interpolation Methods for Spatiotemporal Geographic Data. *Journal of Computers, Environment, and Urban Systems*, 28(3):201-227, 2004.
- [17] P. Revesz. *Introduction to Constraint Databases*, Springer, New York, 2002.
- [18] P. Revesz and L. Li. Constraint-Based Visualization of Spatiotemporal Databases. In M. Sarfraz, editor, *Advances in Geometric Modeling*, pages 263-276. John Wiley Inc., 2003.
- [19] S. Rosenstone. *Forecasting Presidential Elections*, Yale University Press, New Haven, 1983.
- [20] D. A. Shepard. A two-dimensional interpolation function for irregularly spaced data. In *Proc. 23rd National Conference ACM*, pages 517-524, 1968.
- [21] C. Stallings., R. L. Huffman, S. Khorram, and Z. Guo. Linking Gleams and GIS. In *Proc. American Society of Agricultural Engineers*, 1992.
- [22] E. W. Weisstein. Spherical Coordinates. From *MathWorld*—A Wolfram Web Resource.
- [23] S. Wu and P. Revesz. DOAS: A drought online analysis system with constraint databases. In *Proc. of the 4th National Conference on Digital Government Research*, pages 417-418, 2004.
- [24] E. G. Zurfueh. Applications of two-dimensional linear wavelength filtering. *Geophysics*, 32:1015-1035, 1967.
- [25] <http://election.dos.state.fl.us/>
- [26] <http://www.cnn.com/2004/ALLPOLITICS/10/25/florida.poll/>

Scalable Data Collection and Retrieval Infrastructure for Digital Government Applications*

Hanan Samet

Department of Computer Science
Center for Automation Research
Institute for Advanced Computer Studies
University of Maryland at College Park
hjs@cs.umd.edu

Leana Golubchik

CS and EE-S Departments
Integrated Media Systems Center
Information Sciences Institute
University of Southern California
leana@cs.usc.edu

1. INTRODUCTION

In this paper we describe highlights of the project titled “Scalable data collection infrastructure for digital government applications” under the auspices of the Digital Government Research Program of the National Science Foundation. Our research is focused on taking advantage of the distributed nature of data and the interaction with it. Our efforts have been directed at both the systems/theoretical and applications levels. On the systems and theoretical levels, we have continued our development of the BISTRO system (Section 2). On the applications level, work has commenced on the development of a mechanism for spatially tagging text documents for retrieval by search engines based on both content and spatial proximity (Section 3).

2. BISTRO

Hotspots are a major obstacle to achieving scalability in the Internet; they are usually caused by either high demand for some data or high demand for a certain service. At the application layer, hotspot problems have traditionally been dealt with using some combination of increasing capacity, spreading the load over time and/or space, and changing the workload. Previous classes of solutions have been studied in the context of applications using one-to-many, many-to-many, and one-to-one communication. However, to the best of our knowledge there is no existing work, except ours on making applications using many-to-one communication scalable and efficient; existing solutions simply use many independent one-to-one transfers. This corresponds to an important class of applications, whose examples include digital government tasks such as submission of income tax forms to

IRS. We proposed Bistro, a framework for building scalable and secure wide-area digital government upload applications.

Briefly, the Bistro upload architecture works as follows. Given a large number of clients that need to upload their data by a given deadline to a given destination server, the Bistro architecture breaks the upload problem into three steps. Step 1, which is the timestamp step, must be accomplished prior to the deadline for clients to submit their data to the destination server. In this step, each client sends to the server a message digest of their data and in return receives a secure timestamp ticket from the destination server as a receipt indicating that the client made the deadline for data submission. The purpose of this step is to ensure that the client makes the deadline without having to transfer their data which is significantly larger than a message digest and might take a long time to transfer during high loads which are bound to occur around the deadline time. It is also intended to ensure that the client (or an intermediate bistro used in Step 2) does not change their data after receiving the timestamp ticket. All other steps can occur before or after the deadline. Step 2 is the transfer of data from clients to intermediate hosts, termed bistros. This results in a low data transfer response time for clients. Since the bistros are not trusted entities (unlike the destination server), the data is encrypted by the client prior to the transfer. Step 3 is the collection of data by the destination server from the bistros. The destination server determines when and how the data is collected in order to avoid hotspots around the destination server. Once the destination server collects all the data, it can decrypt it, recompute message digests, and verify that no changes were made to a client’s data (either by the client or by one of the intermediate bistros) after the timestamp ticket was issued. A summary of main advantages of this architecture is: (1) hotspots can be eliminated around the server because the transfer of data is decoupled from making of the deadline, (2) clients can receive good performance since they can be dispersed among many bistros and each one can be direct to the best bistro for that client, and (3) the destination server can minimize the amount of time it takes to collect all the data since now it is in control of when and how to do it (i.e., Bistro employs a server pull).

Our main research activities within the Bistro framework have been along the above described three steps. In addition to focusing on performance and security issues, our recent efforts have also included research directions on fault tol-

*This work was supported in part by the US National Science Foundation under Grant EIA-00-91474, as well as the Policy Development and Research Division of the Department of Housing and Urban Development.

erance issues related to the entire Bistro framework. That is, the security mechanisms in the Bistro upload protocols guarantee integrity and privacy of the data being upload. However, to improve the performance characteristics of our scheme, it is still desirable to provide mechanisms and policies for ensuring that data will not have to be retransmitted due to losses or temporary unavailable which could occur due to failures or malicious behavior of various system components.

To this end, our work focuses on augmenting our current Bistro architecture with appropriate fault tolerance and redundancy mechanisms and policies, where the amount of redundancy and degree of fault tolerance depends on the application and the reliability characteristics of the system components. Our goal in this work is to maintain comparable performance to that of a system without fault tolerance mechanisms and to reduce the overhead attributed to fault tolerance mechanisms (such as storage and network bandwidth overheads) as much as possible.

Lastly, this year, we have also focused on designing incentive schemes for encouraging (non-malicious and reliable) participation in the infrastructure. We are currently pursuing a reputation based approach to this problem. Reputation is a measure of how trustworthy a bistro has been in the past. It is also indicative of how much of its own resources a bistro had contributed to aiding others in the infrastructure. The higher the reputation of a bistro, the higher preference it would receive in the allocation of the infrastructure's resources. The incentive schemes are needed to encourage bistros to volunteer their resources as well as to incentivize nodes that are currently contributing resources to behave in a reliable and non-malicious manner. (Examples of malicious behavior include corruption of data or reluctance to forward data to an appropriate destination).

3. SPATIAL TAGGING

Spatial data can be found in a multitude of forms and variety. We are currently involved in building automated tools that can automatically identify and extract spatial information from web documents. Our first study aimed at converting structured documents, such as, EXCELL spreadsheets and semi-structured data, such as XML and GML documents, into spatial data using an interactive tool. Subsequently, we built tools for identifying postal addresses in documents and tools for geocoding these postal addresses to points on a road map. The real challenge is to automatically extract, and recognize references to geographical locations in text, pdf or word document which do not have any underlying structure.

Our goal is to build a search engine that retrieves documents where the similarity criterion is not based solely on exact match of elements of the query string but instead also based on spatial proximity. For example, the user could search for "Housing Projects" in the vicinity of "College Park, MD". Thus, the search has a content and location specifier associated with it. The results would only return such documents that qualify both the content and location specifier that was provided to the system by the user. Our testbed application domain is a set of documents on a website of the Department of Housing and Urban Development with whom we are collaborating on this project. Below, we report some of the progress made in this direction this year.

We started by investigating into algorithms that automati-

cally identify spatial references in text, pdf, word and other unstructured documents. On identifying spatial references in a document, we associate the document with a set of spatial tags. For example, a document that relates to events in *College Park, Maryland*, is assigned a spatial tag corresponding to the latitude/longitude of College Park.

The document tagger makes use of the GNIS dataset which is a publicly available gazetteer containing the names of places in the world. Given a document, one strategy would be to compare every word in the document with the gazetteer to look for potential matches to records in the GNIS database. However, this process is inefficient. First of all, posing queries to the gazetteer is expensive and should be limited to a few sampled words in the documents. Secondly, a word in the document may match to multiple entries in the gazetteer. For example, a word "York" in the document may correspond to a dozen equally likely entries in the gazetteer. Thirdly, there is no mechanism to avoid *false hits*, i.e., a word "nice" in a document may or may not be a spatial reference to "Nice, France". The tagger that we are building resolves these ambiguities by assigning a relevancy measure to each identified spatial location, the geographical distances between the matches, their offset position in the document, and the size of the document. This model has been shown to perform well in a sample test scenario.

Once the tagger has identified a relevant set of spatial descriptors for a document, we must decide the extent of the tag. In particular if the region has extent such as a county or a road then we must decide whether to tag it with the locations of its starting and ending locations or should we just tag it with its centroid? These issues arise for other types of spatial data as well such as counties, countries, states, etc.

Having developed a document tagger, we need to rank the various locations that are specified in the document. This is important in finding the documents most relevant to a given spatial search string. We are working on the development of a number of different spatial ranking algorithms and will evaluate their effectiveness. We will do this by weighting the spatial references. There are a number of options. One is by frequency. Another is by the extent of the distribution of the references to the spatial search string in the document.

To reinforce the importance of ranking we turn for an example to a search for documents related to Hurricanes. Suppose that we are scanning a news archive. It is not unusual to encounter articles in place A (e.g., Singapore) about a Hurricane in place B (e.g., New Orleans). Clearly, the important spatial location here is New Orleans and not the fact that the article appeared in the Singapore Strait Times newspaper.

Finally, we are working on the development of a method to present results to the user that possibly give an indication of the location of the documents as well as the range of the locations referenced by the relevant documents. Alternatively, we may want to rank a collection of documents by the most relevant spatial locations that they reference. For this we are investigating use of the SAND spatial browser developed by our research group. We also plan to try to show users the distribution of the locations referenced by a collection of documents. Until now the SAND spatial browser has been used primarily to respond to spatial queries involving nearest neighbors and ranges. So, this work represents a significant conceptual change in its structure.

Automatic Alignment of Vector Data and Orthoimagery for The National Map

Craig A. Knoblock and
Cyrus Shahabi
University of Southern
California
4676 Admiralty Way
Marina del Rey, CA 90292
{knoblock|shahabi}@usc.edu

Ching-Chien Chen
Geosemble Technologies
2041 Rosecrans Ave., Suite
245
El Segundo, CA 90245
jchen@geosemble.com

E. Lynn Usery
U.S. Geological Survey
1400 Independence Road
Rolla, MO 65401
usery@usgs.gov

ABSTRACT

A general problem in combining road vector data with orthoimagery from different sources is that they rarely align. There are a variety of causes to this problem, but the most common one is that the latest products are collected with higher accuracy and improved processing techniques. In previous work, we developed techniques to automatically correct the alignment of vector data with orthoimagery using a technique called conflation. However, in applying our technique to real-world datasets provided by USGS, we discovered that these techniques failed in some areas. In this paper, we describe some refinements to our original approach that provide consistently better results in aligning the vector data with the orthoimagery.

Categories and Subject Descriptors

I.4 [Image Processing]: Feature Measurement; I.2 [Artificial Intelligence]: Learning

General Terms

Algorithms

Keywords

orthoimagery, vector data, conflation, alignment

1. INTRODUCTION

The *National Map* is a government effort to make geospatial data available for the US beginning with the 133 urban areas of the Homeland Security Infrastructure Program (HSIP). The purpose of this project is to make these integrated datasets available to government organizations to support science, crisis response and emergency planning, among other applications. Currently, the U.S. Geological Survey (USGS) is collecting high resolution 0.3 meter orthoimages under contracts to industry. Vector data including transportation, hydrography, boundaries, and structure



Figure 1: Poor Alignment of Roads with Orthoimagery Before Processing

outlines, from a variety of sources including federal, state, local and tribal governments, must be aligned with the orthoimages (Figure 1). The problem is that there are no automated techniques for aligning vector data with orthoimagery and this is a very labor intensive task.

Under our NSF-funded ITR grant, we developed an approach to automatically align road vector data with high resolution orthoimagery. This approach exploits the road vector data to perform a highly focused search for the corresponding intersection points in the orthoimagery. The result is a set of accurately identified intersections that can serve as control points to align the vector data with orthoimagery. While these techniques provide significantly better alignment of the vector data with the orthoimagery in most places, there are, however, some regions where the alignment of the vector data is worse than the original.

2. PREVIOUS WORK

In previous work, we developed techniques for automatic conflation of road vector data with orthoimagery. The most effective technique we found exploits a combination of the knowledge of the road network with image processing in a technique that we call *localized template matching* [2]. With this approach, we first train the system on a small area of the orthoimagery to learn the road color distribution based on the image pixels' hue value. We then classify image pixels as road/non-road regions by applying a Bayes classifier with this hue distribution. Meanwhile, the system locates road intersection points from the road vector dataset. For



Figure 2: Further Degraded Alignment Caused by the Failure to Locate the Intersections

each intersection point, a template inferred from the vector information (e.g., road width and directions) is matched against the localized area around the intersection to find the corresponding intersection in the pre-classified image. By exploiting the road direction and width information we improve both the accuracy and efficiency of detecting intersections in the image.

An issue that arises is that the localized image processing may still identify incorrect intersection points, which introduces noise into the set of control point pairs. It is essential to use a filter to eliminate misidentified intersections and only keep the accurately identified intersections, hence improving the precision at the cost of reducing recall. We use the Vector Median Filter (VMF), which works based on the fact that there is a significant amount of regularity in terms of the relative positions of the intersections on the vector and the corresponding intersections on the orthoimagery across data sets. More precisely, VMF first interprets the coordinate displacement between the intersections on the vector and corresponding orthoimagery intersections as 2D vectors (termed as control-point vectors). Next, a given intersection is kept as a control point if it is within $k\%$ (k is a predefined constant) of the vectors that are closer to the median vector.

3. IMPROVING THE ALIGNMENT

With the test sets described in [2], the approach described above produced an accurate alignment of the vector data with the orthoimagery. However, in applying our technique to real-world datasets provided by USGS, we discovered these techniques failed in some areas (Figure 2). Two major issues cause this failure: first, some image pixels are misclassified mainly because only one color channel (i.e., hue) is utilized to categorize the road/non-road regions. Second, we set a very high threshold for filtering control points, which provided high precision, but low recall on the control point pairs. We improved the accuracy of the vector-imagery alignment by developing techniques to address these issues.

To improve the road classification, we switched to a machine learning classifier, called a Support Vector Machine (SVM), to categorize image pixels based on all available image color information (i.e., RGB). An SVM maps all training data into a high-dimensional Hilbert space and then generates region boundaries as hyperplanes separating data points. We utilized the freely available SVM library SVMLIB [1] as the



Figure 3: Accurate Alignment of Roads with Orthoimagery After Algorithm Improvements

learning method to learn the road color (RGB) distribution to predict image pixels as road/non-road pixels.

To address the low recall due to the VMF filtering technique, we developed a cluster-based approach to dynamically choose the filtering threshold, k , for diverse regions. In VMF, we used vector median whose summed distance to other neighboring control-point vectors is minimal to filter outliers. The similar vectors tend to form clusters around the median vectors. Based on this observation, we modified the filtering technique to accommodate more vectors that are close to the median vector. More precisely, instead of filtering based on a fixed percentage, we changed the system to dynamically choose different k for diverse areas based on two types of distributions of the control-point vectors: (1) if the vectors form a cluster around the median vector, the system keeps the control-point vectors in the cluster, and (2) if there is no obvious cluster for the control-point vectors, the system will set k to 70%.

4. DISCUSSION

Experimental results show that the new classification technique and improved filtering provides significant improvement on precision and recall for identifying intersections. This in turn results in improved vector-imagery conflation results (Figure 3). The remaining challenge is to improve the accuracy of the intersection detection in areas around highways. Highways pose a particular challenge because of their varying widths. Once we have addressed this issue, we will run a comprehensive evaluation of the techniques in several different regions covered by *The National Map*.

5. ACKNOWLEDGMENTS

This research is based upon work supported in part by the National Science Foundation under Award No. IIS-0324955 under a supplemental grant from the Digital Government Program, and in part by the U.S. Geological Survey under order number 05CRSA0551.

6. REFERENCES

- [1] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [2] C.-C. Chen, C. A. Knoblock, and C. Shahabi. Automatically conflating road vector data with orthoimagery. *To appear in Geoinformatica*, 2006.

National Large-Scale Urban True Orthophoto Mapping and Its Standard Initiative

Guoqing Zhou and Wenhan Xie
Old Dominion University,
Tel: (757) 683-3619
Email: gzhou@odu.edu

Susan Benjamin
U.S. Geological Survey
Western Geographic Science
Center

Robin G. Fegeas
U.S. Geological Survey
Mapping Applications Center-
Reston, Virginia

John Simmers
Virginia Department of
Transportation

Hap Cluff
Dept. of Information
Technology
City of Norfolk

Y. Lei
Dept. of Information
Technology
City of Virginia Beach

Jeanne Foust
Environmental
Systems Research
Institute (ESRI) Inc.

1. CURRENT PROJECT ACTIVITIES

This document would highlight the current project activities, published or unpublished research contributions, success and challenges from March 2005 through December 2005 and plane for the coming years on the project, entitled "National Large-Scale Urban True Orthophoto Mapping and its Standard Initiative" funded by the National Science Foundation (NSF).

There are several ideas in the outcomes of our investigation:

(1) *Polyhedral primitives based CSG R-tree for urban building model.* A kind of model for exactly describing urban 3D buildings for large-scale urban true orthoimage generation is presented. This method is based on CSG (Constructive Solid Geometry), belonging to volumetric representation in computer graphics. This method is well suited to describing complex shapes, which can be composed by a set of primitives. Within this proposed approach the buildings are described by combining a set of basic primitives, such as box, wedge, and rectangular pyramid. This representation model is particularly useful for urban true orthoimage generation because a complex building in this model can be partitioned into many simple building parts, each of them corresponding to a basic building model. Moreover, the model is able to consider diverse building, e.g. flat roof, gable roof, and hip roof building. We implemented this work using 2D plane information according to digital surface model. The 2D plane of a building is divided in rectangles, arcs, and circles, each of the primitives representing the ground plane of a building part. The primitives are combined by means of the Boolean operators: union, intersection, and difference. So, the buildings can be described as a CSG tree, where the leaves contain primitives and the internal nodes contain Boolean operations.

(2) *Accuracy improvement of urban true orthoimage generation.* When applying the existing true orthorectification methods to orthorectify urban high buildings, there exist many problems such as incomplete orthorectification of boundaries of artificial buildings, edge blurring and distortion; incomplete

orthorectification to small buildings in building roofs, and/or balcony due to the lack of proper 3D model representation. Therefore, a novel orthorectification method that is based on 3D R-tree urban building model is presented. This method can exactly solve the problem mentioned above and improve the accuracy of true orthoimage. Combining with the technique of CSG model and LOD (Level of Detail), 3D R-tree can more effectively organize and accurately describe the structure of urban building model. Thus, we can produce the dynamic 3D digital building model (DBM) with varying levels of detail, which ranges from tiny parts of one building to the entire block of street. With the proposed method, automatic extraction of the roof edges from DSM (Digital Surface Model) is first carried out. With different types of roof and height of the building, we can construct the primitives of CSG model. Second, with the spatial properties of minimum rectangular bounding box of primitives, we can build the 3D R-tree object-oriented structure for every building. When back-projecting the 3D R-tree-based DBM onto the original image, preliminary outlines of buildings with vector feature in the original image can be obtained. Thus, accuracy of orthorectification with the geometric constrains of building vector boundaries can effectively be improved. A test field located in downtown of Denver, Colorado has been used to test our methods. The experimental results demonstrated that the proposed method in this paper can effectively improve the accuracy of urban buildings with 3-5 pixels, especially for the boundaries of building, and some small buildings can completely orthorectified.

(3) *Self-calibration of camera for orthorectification.* Because the orthorectification needs the interior elements of camera, which is from camera calibration. Thus, the initial value of orientation parameter in bundle adjustment is very important for the accuracy improvement of orthoimage generation. For large-scale urban aerial images, there exist abundant buildings with line outlines. According to these geometric features, a calibrating method based on vanishing point of multi-view is presented. Without any prior control information, this method can accurately calibrate the interior and exterior orientation parameters of aerial images.

2. PUBLISHED OR UNPUBLISHED RESEARCH CONTRIBUTIONS

- [1] Zhou, G., Chen, W., and Kelmelis, J. A Comprehensive Study on Urban True Orthorectification, *IEEE Tran. on Geoscience and Remote Sensing*, vol. 43, no. 9, pp. 2138-2147.
- [2] Zhou, G., and J. Chen, Urban Large-scale Orthoimage Standard for National Orthophoto Program, *IEEE Geoscience and Remote Sensing Annual Conference*, Seoul, Korea, July 25-29, 2005.
- [3] Zhou, G., National Large-scale Urban TRUE Orthophoto Mapping and Its Standard Initiative, The National Conference on Digital Government Research (Abstract), May 15-18, 2005, Atlanta, GA, USA.
- [4] Xie, W., Zhou, G., Urban 3D Building Model Applied to True Orthoimage Generation, *1st EARSeL Workshop on Urban Remote Sensing*, Berlin, March 2 - 3, 2006.
- [5] Zhou, G., Xie, W., 3D Building Model and True-Texture Reconstruction for Urban Micro-Environment Analysis, to be published to *IEEE IGARSS*, Denver, August, 2006.
- [6] Xie, W., Zhou, G., Camera Calibration with Long Focal Length, to be published to *IEEE IGARSS*, Denver, August, 2006.
- [7] Xie, W., Zhou, G., Accurate Pose and Location Estimation of Digital Camera in Building Scene, submitted to *IEEE IGARSS*, Denver, August, 2006.
- [8] Zhou, G. and W. Xie, Accuracy Improvement of Urban True Orthoimage Generation Using 3D R-tree-based Urban Model, *The 7th Annual International Conference on Digital Government Research*, San Diego, California, May 21-24, 2006.
- [9] Zhou, G. and W. Xie, Accuracy Improvement of Urban True Orthoimage Generation Using 3D R-tree-based Urban Model, to be submitted to *IEEE Tran. on Geoscience and Remote Sensing* in March.

3. IDENTIFIED SUCCESSES AND CHALLENGES

We have successfully developed an integrated system of true-orthophoto generation. This software system is programmed by Visual C++, OpenGL and VRML. So far, the significant technique contributions in this developed system mainly include:

- (1) Representation of urban 3D building model including simplification of TIN terrain model and three-level CSG data structure etc. Traditional raster digital surface model exists plenty of redundant data. The proposed simplification algorithm in our system can greatly reduce the redundancy and improve the operational efficiency.
- (2) Automatic generation of urban model, including artificial object extraction, category and modeling etc. the entire process is a bottom-up procedure from low-level data to high-level knowledge. With the 2D vector

map of building outlines and the building height deprived from digital surface model, model selection, extraction and combination can be automatically implemented.

- (3) Effective organization and management of urban 3D model database. All the object models are organized using topological tree structure. The models described with this structure can be easily retrieved and decomposed.
- (4) Accurate generation of true orthorectification, including CSG model-based matching and rectification. The proposed method for true orthoimage generation on the basis of CSG model is more accurate.
- (5) 3D visibility analysis, including occlusion detection and street visibility. In a true orthoimage, buildings should be presented in their true upright planimetric positions. However, the walls in original images occluded streets or other objects. Thus, occlusion detection should be analyzed. And occlusion compensation is implemented by refilling the occluded areas from neighbor slave orthoimages;
- (6) Flexible operation of 3D true orthophoto, such as pan, zoom, flythrough, query and statistics.

4. PLANS FOR THE COMING YEAR

This is a two-year project. In the first project year, we have achieved significant progress. In this second project, we plan to deploy the following activities.

- 1. True texture mapping to generate urban orthophotomap. For some applications, such as urban microclimate investigation, texture of building wall is needed. Urban orthophotomap will largely extend the applications of true orthoimage in urban planning, environment analysis, especially micro-environment analysis.
- 2. Further development of true orthophoto system. The system developed in the first project year need to be refined in the second project year, and then the refined system will be tested and evaluated via several co-PI, Co-I and partners, such as City Planning Office at City of Virginia Beach, Virginia Department of Transportation.
- 3. Evaluation of the existing national orthophoto accuracy and comparison with our methods.
- 4. Standard initiative of orthophoto in urban area in collaboration with all Co-PIs, Co-Is and partners.
- 5. Workshop on urban orthophoto, or publishing special issue on urban orthophotos for discussing urban orthophoto standard problems.

SESSION 8A

PROCESS AND WORKFLOW

Moderator

Lois Delcambre, Portland State University, USA

Titles and Authors

Lynx: An Open Architecture for Catalyzing the Deployment of Interactive Digital Government Workflow-Based Systems
Vélez, Iván P.; Vélez, Bienvenido

Argos: Dynamic Composition of Web Services for Goods Movement Analysis and Planning
(Abstract)
Ambite, José Luis; Giuliano, Genevieve; Gordon, Peter; Pan, Qisheng; Jinwala, Mountu;
Kapoor, Dipsy; Wang, LanLan

Data Processing Workflows in the Social Sciences: Representation and Automatic Generation
Ambite, José Luis; Kapoor, Dipsy; Jinwala, Mountu

Lynx: An Open Architecture for Catalyzing the Deployment of Interactive Digital Government Workflow-Based Systems

Iván P. Vélez

Advanced Data Management Group

Electrical and Computer Engineering Department

University of Puerto Rico, Mayagüez Campus

Mayagüez, PR 00681-9042

1-787-831-3244

i_velez@computer.org

Bienvenido Vélez

Advanced Data Management Group

Electrical and Computer Engineering Department

University of Puerto Rico, Mayagüez Campus

Mayagüez, PR 00681-9042

1-787-831-3244

bvelez@acm.org

ABSTRACT

We introduce Lynx, a new email extension for workflow systems based on Web Services. Web service based workflows provide support for aggregating web services into new higher-level web services by means of process composition. This approach does not usually support direct interaction with people. On the other hand, traditional collaboration tools like email or instant messaging do not provide the necessary support for structured business processes. Lynx provides a web service through which a workflow application can interact with human partners via an email based forms interface without requiring a specialized client. We constructed a Lynx prototype and tested it with the ActiveBPEL engine. User interaction is achieved by means of XForms dynamically generated by Java classes dynamically loaded based on the XML schema of the documents exchanged. We illustrate the usefulness of our approach in a Digital Government scenario.

Categories and Subject Descriptors

H.3.5 [Information Storage and Retrieval]: Online Information Services – Web-based services

H.4.1 [Information Systems Applications]: Office Automation – Workflow management

General Terms

Performance, Design, Standardization, Languages

Keywords

BPEL, Digital Government, E-mail, Web Services, Workflow, XForms, XML

1. INTRODUCTION

Digital Government information systems provide support for government officials to satisfy their citizen's needs. Such systems could go from simple information displays, to systems that automatically canalize user requests throughout a government agency, maintain relevant document data and improve the overall quality of the services provided to the citizen. Workflow systems are sometimes adopted to automate government processes by specifying how tasks are structured, who performs them, what their relative order is, how they are synchronized, and how information flows to support the tasks.

Workflow systems allow the specification and evolution of complex business processes without requiring complex programming skills. Such specifications can be automatically compiled to generate actual code that implements the process. Ordinary business process workflows are oriented towards interacting with human users directly via some interface that runs at their workplace desktop [9]. This often requires a custom software client and either physical presence or reliable remote access to the workplace. This approach typically follows either a pull-based model, where the user is burdened with periodically logging in and inspecting the system to verify the status of pending workflow transactions [8] requiring their attention, or only provide support for simple notifications based on various protocols such as email or instant messaging. On the other hand, web-service based business processes provide support for aggregating web services into new higher-level web services by means of process composition [14]. This approach often provides insufficient support for direct or synchronous interaction with people. Collaboration tools like email or instant messaging do not provide the necessary support for structured business processes. Lynx attempts to overcome some of these limitations by enabling email as an alternative mechanism for interaction between a web-service based workflow process and human users.

Email has the potential to free participants from the constraints of space and time allowing senders and recipients to communicate at convenient times and places [16]. Email is familiar for people, provides a simple means of communication for person to person interaction, allows easy interconnectivity between all participants, and most importantly it allows mobility and capability of working

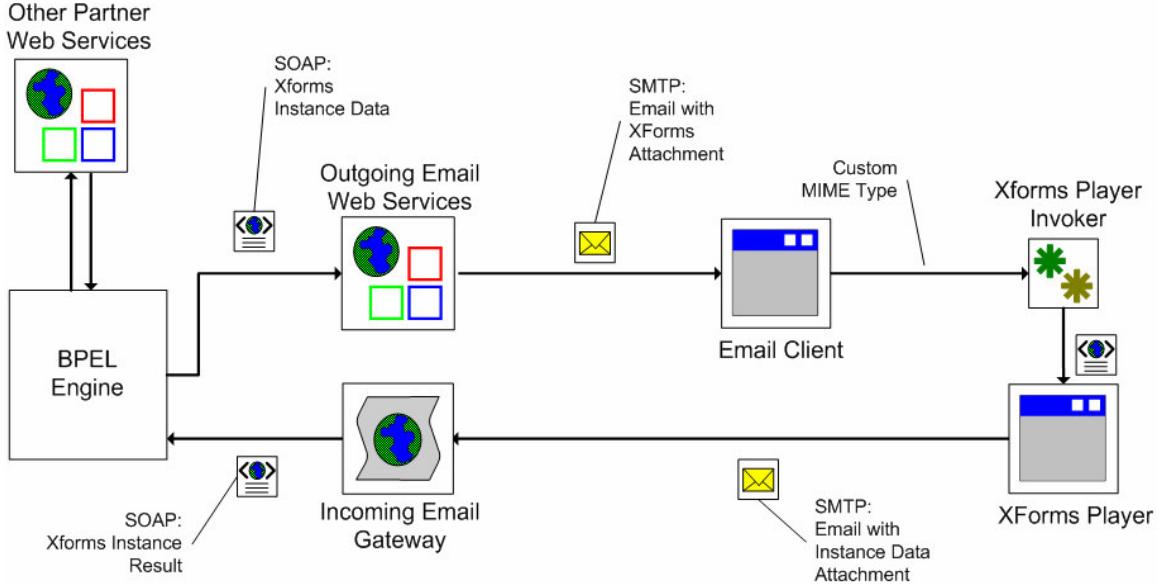


Figure 1. Lynx Architecture

from a distant location through the Internet. Our experience in regional and municipal Digital Government environments has consistently demonstrated the need for familiar and broadly accessible interaction mechanisms for the effective adoption of information technologies (IT) given the relatively low level of exposure of personnel to IT in these settings. Furthermore, from a technical standpoint, email operates using simple ubiquitous protocols available across links of widely varying qualities, firewalls and other security membranes.

Lynx can extend many web service based workflow engines with the ability to send transaction requests to human workflow partners using email not only as the transport mechanism, but rather as the interaction application. Using Lynx users carry out workflow transactions by processing electronic forms transported to/from their email accounts. The current prototype supports the XForms [12] standard and generates each form based on the XML schema of the document being transported. The request for the transaction is generated by a business processes typically specified using a language such as the Business Process Execution Language (BPEL). BPEL defines a notation for specifying business process behavior based on web services [6].

This paper describes the Lynx prototype and its architectural integration into an open web service based email-enabled workflow system. The paper illustrates the usefulness of this system in a Digital Government scenario.

The remainder of this paper is organized as follows. Section 2 discusses the Lynx's architecture. Section 3 illustrates its application to digital government. The integration of the system with the workflow process will be described in Section 4. Section 5 describes some of the previous research work related to communication frameworks and business processes. Finally, Section 6 presents our conclusions and suggests some areas for future research.

2. ARCHITECHTURE OVERVIEW

Figure 1 shows the different elements that comprise a workflow architecture integrating Lynx. The server side is composed of a BPEL execution engine, an outgoing email web service and other partner web services, and an incoming email gateway. The workflow engine exports a web service interface that can be used to initiate and interact with a business process. Each business process running in the BPEL execution engine may interact with multiple business partners exporting their own WSDL web service interfaces. Lynx consists of two modules: an outgoing email web service and an incoming email gateway. The client side is composed of a standard email client application and an XForms player component. The following subsections describe the role of each of these components.

2.1 BPEL Execution Engine

The BPEL Execution Engine provides the workflow management capabilities. BPEL processes executed by this component interact with the external world through web services [14]. The workflow process is specified in an XML-based language. BPEL defines a model and grammar that describes the behavior of a business process based on interactions between the process and its partners. The interaction with each partner occurs through web service interfaces. The BPEL process defines how multiple concurrent service requests from these partners are coordinated to achieve a business goal, as well as the state and the logic necessary for this coordination. By composing services into new, more complex web services, BPEL allows creation of heterogeneous distributed applications [15]. The BPEL engine can run business processes for hours, days or months, and may invoke other long-running services. A BPEL process can contain steps that require waiting for external events or human interaction by invoking a web service that handles this type of interaction. In this case, the BPEL process can invoke the Lynx web service as described in the following section.

2.2 Outgoing Email Web Service

Lynx's outgoing email web service provides the necessary services to interact with human partners through email. It dynamically generates the email message containing the document sent when a process needs to interact with a human partner. The service accepts documents to be processed by a human partner via its web service interface. In response the service automatically generates an electronic form for the document and sends it as an attachment to the human partner via email.

2.3 Partner Web Services

Other optional partner web services may provide other services required by the BPEL processes. These services can include document validation, external notifications, transaction logging, document storage in an external database, and other external processes such as transactions that need to be completed by a business process of another government agency.

2.4 Incoming Email Gateway

An email server is periodically monitored by the Incoming Email Gateway that listens for incoming email messages generated by interactions with human partners. It forwards any received processed documents to the appropriate step within a running BPEL process thus allowing it to continue its workflow.

2.5 Email Client

Any standard email client can be used to receive emails. The emails received by the users will have an attached document that can be viewed with the XForms player component. The MIME type of the attached document is defined as a custom application/type registered in the client to be able to view it with the corresponding XForms player.

2.6 XForms Player

This component acts as a plug-in that renders the document received through email as an electronic form with controls that allow more sophisticated interactions than HTML forms. XForms allow data to be validated by the browser, such as types of fields being filled in, that a particular field is required, or that one date is later than another. XForms are also device independent, meaning that the same form can be delivered without change to a traditional browser, a PDA, a mobile phone, a voice browser, and even an email client. XForms are themselves XML documents and can be filled from other XML documents called instance data. All these features of XForms [13] have the potential for dramatically reducing the amount of custom GUI code necessary to implement client applications and improving the user experience by giving immediate feedback of what is being filled in. We hypothesize that by exploiting XForms we will demonstrate the viability of implementing complex distributed interactive applications with significantly less coding. Additionally, simple modifications to the interaction screens will often not require re-programming of GUI code. This is of particular importance in Digital Government environments where programming skill are severely scarce and difficult to hire.

3. A DIGITAL GOVERNMENT SCENARIO

This section describes a sample Digital Government scenario where our proposed architecture is expected to be adopted. The Puerto Rico Registry of Deeds holds a public archive that contains

all the documents about property transactions, wills, judicial orders and other legal documents. The Registry of Deeds has a large backlog of documents pending for processing due to the meticulous verifications and validations that are currently manually conducted by several specialized human analysts. Many of these processing steps can be automated. However, the required technical expertise is often not easily available. Our hope is that a workflow system based on Lynx will significantly increase document throughput while simultaneously reducing cost and increasing reliability and quality of service at a cost more affordable by small and regional governments.

Figure 2 illustrates a simplified version of the process of registering the purchase of a property in Puerto Rico. This process is similar in many other parts of the world. The notary public, a lawyer in Puerto Rico's system, writes a property title deed document and submits it along with a summary called a presentation minute. This document is received by a receptionist that makes some initial validations.

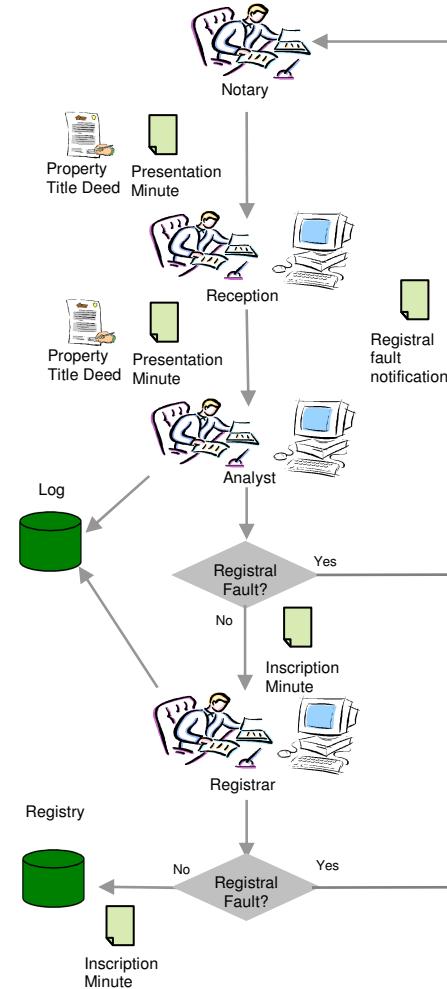


Figure 2. Registry of Deeds Business Process

Then, the document is passed to a suitable analyst to verify that the information contained in the document is correct and matches with the documents already in the Registry pertaining to previous

transactions on the same property. The document can be verified by several analysts if necessary. The analysts generate a second summary of the document, called an inscription minute, and can add annotations if there are minor errors that require clarifications or explanations. If there is a serious incongruence (a.k.a. a registral fault), the document is returned to the notary public. After validation, the document reaches the Registrar himself. This person certifies that all the information is correct and officially adds the inscription minute to the Registry.

The entire process may take several days or weeks through which the workflow system must keep track of every step and provide a query interface to find out the status of every running process at all times. This scenario is typical of many other governmental processes that require a long-running sequence of validations and approvals involving multiple people and systems making decisions and providing information.

4. WORKFLOW PROCESS USING LYNX

This section outlines the integration of Lynx in the specification of the sample business process from Figure 2. The integration of Lynx in a workflow process involves a process reengineering. This means that the processes by which the organization, the Registry of Deeds in this case, creates value and does work should be thought again and redesigned, ridding it of operations that have become antiquated. The process reengineering we propose involves the receipt of the documents in electronic format instead of in paper. In the Registry of Deed's actual computer-based system employees have to manually fill in the information from manuscript documents they receive at their offices, or from document submissions they receive via fax. Instead of increasing the throughput of transactions of the process, this current automation places more responsibilities on the reception and analyst employees. Conversely, if the documents were received in an electronic format, the complete process could be automated beginning in the notary public offices where the documents originate. This would make all the required information available from the start and will result in a more agile business processes.

4.1 Document Definition

To be able to incorporate Lynx to a web services based workflow, the document type XML schemas must first be defined. Document schemas consist of several pre-designed schema templates based on actual Registry of Deeds inscription minutes. These templates include common information that is used among different types of documents. For example, many Registry of Deeds inscription minute documents have the following features that are common with the over 160 different documents used in the Registry of Deeds:

- Header – property number, town, inscription number, description, obligations, and title holders
- General information – principal, interest, affidavit number, appraisal, rights, expiration, descriptions
- Entities – the person or corporation that sells or buys involved in a transaction
- Presentation information – date and time presented, seat number, journal number, town
- Annotation, Log and Attachment information needed by our prototype

These definitions common to all the documents are placed in a master XML schema file that can be imported into any of the specialized XML schemas that implement the definition of the Registry of Deeds documents. The imported schema is then assigned a namespace <http://ece.uprm.edu/RegPropCommon> that identifies the set of elements and attributes of the data types to be used in another document. Accordingly, these schemas must be also imported into the BPEL process definition.

4.2 Process Description

Our current prototype executes the workflow of our sample scenario using the ActiveBPEL [2] implementation of BPEL. The BPEL process is deployed as a message-style web service within an Apache Axis server used by the ActiveBPEL engine. Message-style web services are used because they allow the creation of more document-centric interactions and allow the data to be expressed more naturally, when compared to XML-RPC. Also, message-style directly uses industry standard schemas, provides maximum power and extensibility, avoids building upon assumptions about implementation platform, and allows flexible mapping of platform data structures to XML. This lets the web services work without the need to use any Java-XML binding that would require SOAP encoding, and JavaBeans for each type of document. This approach enables us to have a single web service with a generic operation instead of many web services or a web service with many overloaded operations, one for each different type of emailed document.

The outgoing email web service is made generic by having a single Java method processing any XML that arrives at the service by using a message-style provider. Therefore, the web service would not need to be recompiled nor redeployed in case that a new type of document is sent to the outgoing email web service. Furthermore, our system is made extensible by allowing the addition of new document types to be included in the workflow and be accepted by the web services. The only changes required are the BPEL description, including the processing of a new data type, and adding importing of the new data type's XML schemas [19].

```
<xsd:schema xmlns="http://www.w3.org/2001/XMLSchema"
targetNamespace="http://registro.egov.ece.uprm.edu">
  <xsd:complexType name="EmailInfoType">
    <xsd:all>
      <xsd:element name="to" type="xsd:string"/>
      <xsd:element name="from" type="xsd:string"/>
      <xsd:element name="subject" type="xsd:string"/>
      <xsd:element name="callback" type="xsd:string"/>
      <xsd:element name="class" type="xsd:string"/>
      <xsd:element name="body" type="xsd:anyType"/>
    </xsd:all>
  </xsd:complexType>
  <element name="EmailInfo" type="tns:EmailInfoType"/>
</xsd:schema>
```

Figure 3. XML Schema for Outgoing Email Web Service

The BPEL process invokes the Lynx web service by encapsulating the information necessary to communicate with a human partner inside a message that follows the *EmailInfo* schema depicted in Figure 3. The schema contains the destination human partner email address, subject, callback information, the name of the Java class that implements the appropriate XForms, and the specific document XML instance data payload. To achieve a generic web service the document payload is defined as a schema element of standard XML type *anyType*, thus accepting any document type.

We show in Figure 4 an excerpt of the BPEL specification preparing and performing the web service invocation that will ask the human analyst to validate the documents and return it with any annotations, recommendations or attachments. The invoke operation only has one input variable, and has no output because the interaction with the Analyst was specified asynchronous. This is a design choice for the business process that is not required by Lynx. Nevertheless, asynchronous invocation is favored since an excessive wait time could cause a timeout in the BPEL execution engine's Axis web service operation. Figure 5 shows the main features of a Lynx BPEL process: the variable, correlation and partner definitions, and several parts of the process.

```

<assign name="AnalystAssign">
<copy>
<from expression="egov@ece.uprm.edu" />
<to variable="emailTempInfo" part="EmailInfo"
query="/EmailInfo/to" />
</copy>
. .
<copy>
<from expression="CancelacionHipotecaDirectaAnalyst" />
<to variable="emailTempInfo" part="EmailInfo"
query="/EmailInfo/callback" />
</copy>
<copy>
<from expression=
"edu.uprm.ece.egov.registro.xforms.AnalystXForm" />
<to variable="emailTempInfo" part="EmailInfo"
query="/EmailInfo/class" />
</copy>
<copy>
<from variable="inputDocument"
part="CancelacionHipotecaDirecta" />
<to variable="emailTempInfo" part="EmailInfo"
query="/EmailInfo/body" />
</copy>
. .
</assign>
<invoke name="SendMail" partnerLink="email"
portType="eml:SendMail" operation="sendEmail"
inputVariable="emailInfo" />

```

Figure 4. Invoking an Analyst in BPEL

Lynx's outgoing web service accepts arbitrary XML documents of type *EmailInfo* inside the body of received SOAP messages. This web service then extracts the encapsulated information from the *body* element, and uses it to construct an email message. The web service also needs to generate an XForms document specific to the type of document received and the type of interaction desired from the human partner. The XForms document can provide multiple views of the same instance data. Lynx supports this flexibility by implementing a Java interface (*XFormsInterface*) that includes a method that generates the desired form. The class that implements this interface is specified in the *callback* element of the *EmailInfo* message. This approach has the advantage of allowing the business process to choose the appropriate view for a particular transaction within a process. Lynx dynamically loads the class that implements the interface, instantiates it, and then invokes the method to create the specific XForm using the *body* part of the *EmailInfo* message as instance data. This newly generated XForms is sent as an attachment, with a custom MIME type, to the email address specified in the *To* part of the *EmailInfo* message.

The outgoing email web service also keeps track of the specific correlation information of a document. It specifies this information in the subject of the email message sent to the human partners. This feature helps the email clients organize received email messages by thread so the human partners can easily locate

```

<?xml version="1.0" encoding="UTF-8"?>
<process xmlns="http://schemas.xmlsoap.org/ws/2003/03/business-process/" xmlns:bpws="http://schemas.xmlsoap.org/ws/2003/03/business-process/" name="RegistroProcess"
targetNamespace="http://registro.egov.ece.uprm.edu"
xmlns:tns="http://registro.egov.ece.uprm.edu"

xmlns:xsd="http://www.w3.org/2001/XMLSchema">
<variables>
<variable name="inputDocument"
messageType="tns:cancelacionHipoteca" />
<variable name="emailInfo" messageType="tns:emailInfo" />
<!-- different variables for each operation... -->
<variable name="inputDocumentEmailResponse"
messageType="tns:cancelacionHipotecaAnalyst" />
<variable name="inputDocumentEmailRegistradorResponse"
messageType="tns:cancelacionHipotecaRegistrador" />
. .
</variables>

<correlationSets>
<correlationSet name="registroCorrelation"
properties="tns:ID" />
. .
</correlationSets>

<partnerLinks>
<partnerLink name="service" partnerLinkType="tns:EGovTestPLT"
myRole="service" />
. .
<partnerLink name="email" partnerLinkType="tns:Email"
partnerRole="email" />
</partnerLinks>

<partners>
<partner name="service">
<partnerLink name="service" />
</partner>
. .
<partner name="email">
<partnerLink name="email" />
</partner>
</partners>

<pick createInstance="yes" />
<onMessage partnerLink="service"
portType="tns:EGovTestPT"
operation="setDocument"
variable="inputDocument" />

<correlations>
<correlation set="registroCorrelation" initiate="yes" />
</correlations>

<sequence>
<assign>
<!-- Encapsulate document in email message
As depicted in Figure 4. . . -->
</assign>
<invoke name="SendMail"
partnerLink="email" portType="eml:SendMail"
operation="sendEmail" inputVariable="emailInfo" />
<!-- Process takes decisions depending on email reply... -->
<pick createInstance="no" />
<onMessage partnerLink="service"
portType="tns:EGovTestPT"
operation="analystCompleted"
variable="inputDocumentEmailResponse">
<correlations> . . . </correlations>
<empty/>
</onMessage>

<onMessage partnerLink="service"
portType="tns:EGovTestPT"
operation="reportDocumentError"
variable="notificationFromEmail">
<correlations> . . . </correlations>
<sequence name="AnalystReturnSequence">
<assign name="ReturnAssign"> . . . </assign>
<!-- Send a notification of Failure -->
<invoke name="ReturnNotification" . . . />
<terminate/>
</sequence>
</onMessage>
</pick>
<!-- Here goes other workflow for other document types... -->
</pick>
</process>

```

Figure 5. BPEL Structure of a Lynx process

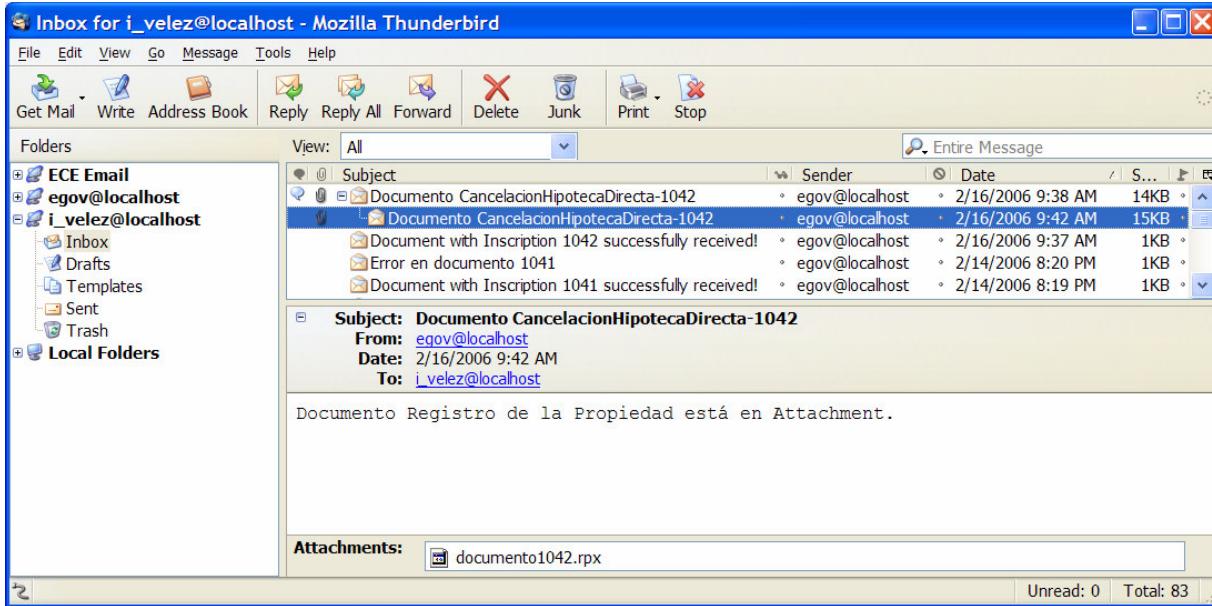


Figure 6. Email Client showing threaded Lynx messages

all the documents referring to a specific case and/or document type as illustrated in Figure 6.

At the client end, the XForms player component launches when an email message containing an XForms document attachment is received and the user opens the attached document. The email client knows it is a Lynx document that needs to be viewed in the XForms player due to the attachment's custom MIME application/type. In our current prototype, we use Chiba [7], an Open Source Java implementation of the W3C XForms standard. Chiba implements XForms by rendering them as standard HTML. The Chiba servlet is a component that runs on a servlet container such as Jakarta Tomcat [4]. Therefore, Tomcat must be running locally on the client computer to support the XForms rendering. Although Chiba is a servlet-based implementation, it implements the whole XForms standard unlike most current web browsers, and works with every browser unlike other client-side XForms implementations. In the future, when mainstream email and browser applications adopt and implement the complete XForms standard, the XForms player component using Chiba could be replaced with the email or web browser client, or a native plug-in that supports the whole standard. This will reduce the applications required at the client to a web browser or an email client.

The location of the XForms attachment is subsequently passed to a web browser that submits it to the Chiba servlet for rendering. Figure 9 shows a sample XForms implemented in the prototype for validating and submitting an inscription minute that the Analyst must validate.

The Analyst returns the document back to the server side after completing, validating, annotating or adding any necessary attachments to the document received by pressing a submit button. This submission element defined in the XForms has an action pointing to a mailto URL (see Figure 7). In response, Chiba submits the updated XML instance data of the corresponding document together with all attachments and

```
<xf:submission
    action="mailto:receiver@ece.uprm.edu?server=
ece.uprm.edu&sender=sender@ece.uprm.edu&subject=Property Registry Document"
    id="s01" replace="none" method="post"/>
```

Figure 7. XForms Email Submission

annotations back to the incoming email gateway on the server side via email.

The Incoming Email Gateway will retrieve the revised document submitted by the Analyst from any standard POP3 email server. All the information necessary to access the incoming email account can be found in a configuration file since the all returned email messages will be addressed to the BPEL engine itself. The incoming email gateway must then invoke a callback web service exported by the business process in order to allow it to continue. This invocation requires producing a SOAP message of the correct operation within the BPEL process. However, what arrives in the email is only the XML instance data. Furthermore, there is no place to specify which operation to invoke in the SOAP message itself since the BPEL process is a message-style service. We solve this problem by dispatching the correct operation based on the type of message.

```
<wsdl:message name="AnalystMessageName">
    <wsdl:part name="CancelacionHipotecaDirectaAnalyst"
    type="CHD:CancelacionHipotecaDirectaType"/>
</wsdl:message>
<wsdl:message name="RegistradorMessageName">
    <wsdl:part name="CancelacionHipotecaDirectaRegistrador"
    type="CHD:CancelacionHipotecaDirectaType"/>
</wsdl:message>
```

Figure 8. WSDL message definitions

We defined different SOAP message part names in the process' WSDL interface definition using the same schema type. This is also done for every variable that is accepted for each different operation in the BPEL process definition. It can be seen in Figure 8 that both messages have the same XML schema type, but they

The screenshot shows a Mozilla Firefox window displaying an XForms document for 'Cancelación Total de Hipoteca a Favor de Persona Determinada (Hipoteca Directa)'. The document is from the 'Estado Libre Asociado de Puerto Rico, Departamento de Justicia, Registro de la Propiedad'. The form includes fields for 'Verificado por' (Ivan Velez), 'Fecha' (10.02.2006), 'Sección' (Mayaguez), 'Número de Finca' (1234567890), and 'Inscripción' (987654321). A calendar dialog is open for February 2006, with the 10th highlighted. Other sections include 'DESCRIPCIÓN' (Descripción conforme con la inscripción 12345), 'CARGAS Y GRAVÁMENES' (Afecta a las cargas que surgen del Registro), 'TITULARES' (Inscrita esta finca a favor de Juan del Pueblo, según surge de la inscripción 12345), and 'TÍTULO QUE SE CANCELA' (Hipoteca a favor de Juan del Pueblo, por la suma de 100,000, según surge de la inscripción 12345). The status bar at the bottom left says 'Done'.

Figure 9. XForms for a Mortgage Document

have different names, depending on which operation they are associated with.

The type of message is specified by the callback part of the *EmailInfo* message that was extracted when the XForms document was created. The message type is used as the root element in the body of the SOAP message returned to the BPEL process. Messages sent to a processes need to be delivered not only to the correct destination web service port, but also to the correct instance of the business process. The process dispatches the message to the appropriate operation within the correct process instance by using the BPEL correlations mechanism. The correlation information serves as an ID for a specific instance of a business process. For example, a social security number might be used to identify an individual taxpayer in a long-running multiparty business process regarding a tax matter. A social security number can appear in many different message types, but in the context of a tax-related process it has a specific significance as a taxpayer ID [6].

In our prototype we use email-based submission and process operation invocation to start a workflow. The initial electronic version of the document produced by the notary public, or a receptionist at the Registry of Deeds that receives a manuscript, is created with empty XForms templates of the documents that are filled with the required information, and then submitted using exactly the same process used by an intermediary such as an analyst. Thus, in Lynx it there is no difference between the initial submission of a document and intermediate submissions made by any other human partner.

In the case of an initial document submission, the Incoming Email Gateway will automatically recognize a new document in the inbox. The document type specifies that the first operation in the BPEL process must be invoked and the BPEL execution engine starts a new instance of the process. If a process with the same correlation information already exists, the submission is not accepted and a response is given by email notifying that a new process could not be instantiated because a document transaction with the same ID is already in progress.

The BPEL operation that waits for a response of the specific type specified in the callback will get invoked automatically. In this way the process resumes at the right point and continues along the process specification. Afterwards, it prepares the next human partner interaction, and sends a request via the outgoing email web service as demonstrated previously in Figure 4. One minor disadvantage of this approach is that it requires any callback web service responding to email-based transactions to use document-style invocations, and declare multiple variables of the same type.

4.3 The Workflow Querying Subsystem

In addition to the main architectural components of Lynx, described in Section 2, our current prototype contemplates additional functionality necessary to build a complete system that can enable the use of a web service based workflow in a government scenario. We are augmenting the basic Lynx web service based workflows that can be constructed with BPEL to support authentication and transaction logging. A web service based workflow has no inherent concept of users. Thus, we store

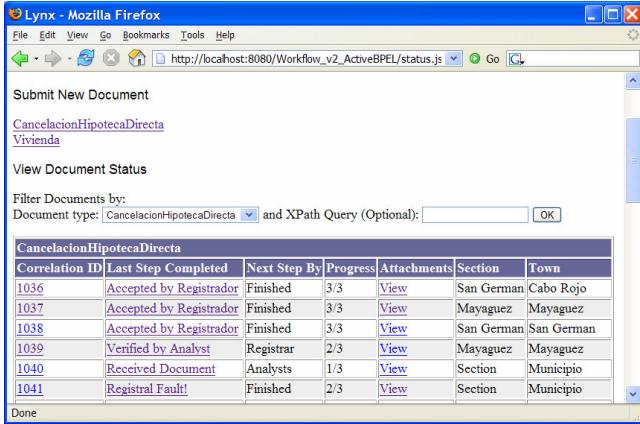


Figure 10. Document Status Page

user information to provide for authentication in a MySQL database. Figure 10 shows the initial prototype of a web-based interface to monitor the status of documents pending processing, and shows documents successfully processed and registered in the Registry of Deeds. Properly authenticated users, such as a supervisor, the registrar, a notary public or an analyst, can inspect the status of a transaction and monitor the progress of a document throughout the business process from within this interface. This transaction logging capability is implemented by having every XML document instance carry its own embedded log that records which steps the specific document instance has gone through. A *LogType* XML schema type imported into every document's schema is showed in Figure 11. The *LogEntryType* complex type includes the date, the process step finished, and who did it. A similar pattern is used for the attachments and annotations in a document. The attachments and a read-only view of the document are accessible as well.

```
<complexType name="LogEntryType">
  <sequence>
    <element name="date" type="date"/>
    <element name="step" type="string"/>
    <element name="by" type="string"/>
  </sequence>
</complexType>
<complexType name="LogType">
  <sequence>
    <element name="entry"
      type="RegPropCommon:LogEntryType"
      minOccurs="0" maxOccurs="unbounded" />
  </sequence>
</complexType>
```

Figure 11. Log XML Schema

Figure 12 and Figure 13 show the log and attachments view, respectively, from the document status web-based interface. Attachments are valuable since they can be used to include important documents needed for the completion of a specific case regarding a document, such as an image file of the property title deed itself, or any other relevant complementary documents.

Furthermore, ActiveBPEL allows the monitoring of the deployed BPEL processes. It provides a web-based administration tool that shows a graphical view of the current state of the workflow, including the process variables and step in the workflow. This is depicted in Figure 14.

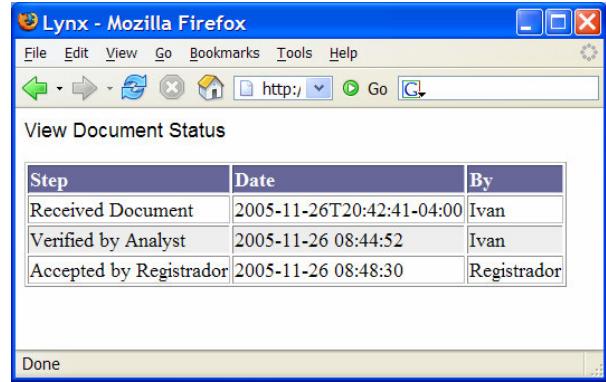


Figure 12. Document Status Log view



Figure 13. Document Status Attachments view

4.4 The XML Storage Subsystem

A persistent copy of all documents is maintained in both a MySQL database that is used by ActiveBPEL for process persistence, and in a Xindice [5] XML database that is used by Lynx to store the current version of each document. The benefit of a native XML Database is that we don't have to worry about mapping our XML documents to some other data structure. While XML documents are organized as tree structures, relational databases organize data in a tabular or grid-like fashion, and use relational linking in order to expose hierarchical constraints on the data. Thus, a lot of flexibility is gained through the semi-structured nature of XML and the schema independent model used by Xindice.

We just insert the data as XML and retrieve it as XML using XPath [20] as the query language and the XUpdate [21] language to insert and update XML documents. This is particularly helpful since we have very complex XML document structures that would be difficult, or impossible, to map to a more structured relational database.

Finally, Lynx is made adaptable to any application requiring human interaction with a web service based workflow by giving the option to customize and configure almost every aspect of its functionality. Any BPEL process can call Lynx's web services to accomplish communication with human partners. The configuration options are kept in XML files so that they can be

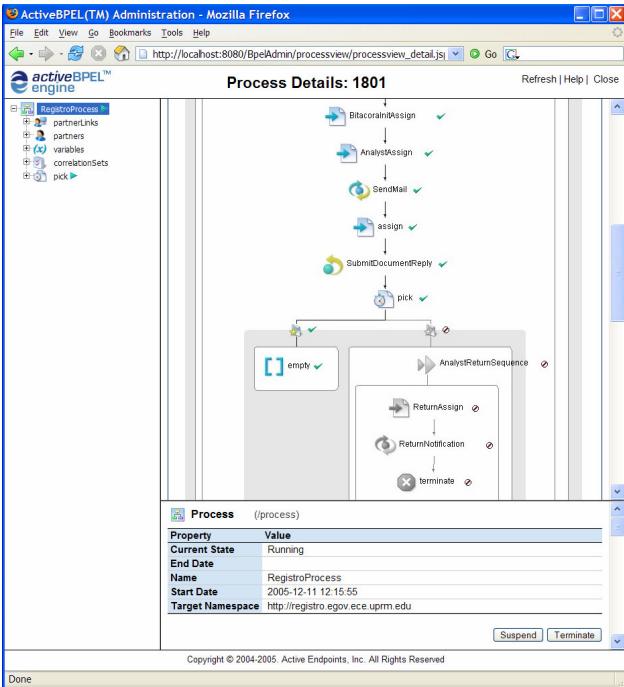


Figure 14. Workflow Process Detail Graph

easily modified with the parameters that suit the specific application. The different types of documents in use are specified along with their correlation information and relevant features (see Figure 15). The correlations are specified as XPath queries of the desired values in the document type's XML Schema. The POP3 and SMTP server addresses to be used by the incoming email gateway and outgoing email web service, respectively, are also specified in other XML configuration files.

```
<config>
  <type>
    <docType>CancelacionHipotecaDirecta</docType>
    <docCorrelation>header/incription</docCorrelation>
    <relevant>
      <Section xpath='header/section' />
      <Town xpath='datosDePresentacion/municipio' />
    </relevant>
  </type>
  <type>
    <docType>Vivienda</docType>
    <docCorrelation>header/incription</docCorrelation>
    <relevant>
      <Notary xpath='DocumentoPresentado/NombreNotario' />
    </relevant>
  </type>
</config>
```

Figure 15. Configuration for Document Types

4.5 Customizing Lynx for Other Scenarios

A similar approach as the one presented for the Registry of Deeds scenario using Lynx could be easily applied to other government applications. The steps needed to integrate Lynx into a web services based workflow application are fairly straightforward. First, the XML schemas for the documents that will be handled by the process must be defined. Then, the WSDL interface for each of the partner web services the BPEL process will invoke, including Lynx's outgoing email web service must be specified. Afterwards, the different operations that the BPEL process will expose must be defined in the WSDL. These operations are

specified by having a different variable for the message that will accept each of the operations, since we use message-style web service operations. Then, the BPEL process that defined the workflow desired should be specified. Also, the XML configuration files specifying the different types of documents, correlation queries, and email servers will need to be customized for the specific application. In addition, the implementation of the XForms for each view for each document must be done, including a Java class that implements the interface to create the XForms for each document. Finally, the partner web services, and the BPEL process itself are deployed. In summary, developing a completely different application will require very little code to be programmed and in particular will dramatically reduce the amount of custom GUI code required.

5. RELATED WORK

Related work can be classified in: pervasive enablement of business processes, middleware architectures, workflow systems, and communication frameworks. In this section we compare some of these systems emphasizing the features that distinguish them from our system.

The Apache Axis Mail Transport [3] allows the transmission of SOAP messages through SMTP mail transport. Its intention is to send web service messages between web services and not between a web service and a human user via email. The receiving application must be capable of extracting the payload from the SOAP message which may require running a web service on the client side. Lynx acts as a gateway between a web services based workflow process and a human partner.

Chakraborty proposes a system called PerCollab which allows convenient communication and collaboration mechanisms (such as SMS, IM and email) to support the activities of a workflow [8]. However, they had to extend IBM's BPWS4J [11] language implementation to implement this functionality. Lynx does not require modifications to the workflow language. It doesn't even require a specific BPEL execution engine.

GreenBSN [22] is a middleware architecture for supporting mobile business service networks. Its main goal is to allow vendors to sell their software as a service, using mobile wireless devices. GreenBSN's service output adaptor module is similar to Lynx because it delivers communication through a user's preferred channel, such as SMS or email. However it is mostly used for asynchronous notifications that only deliver the result of a business process. Lynx allows complete interaction with business processes by allowing human partners to both receive and send information pertinent to the workflow.

Podgayetskaya [17] proposes an architecture and model for business process support for e-government using a workflow engine and web services. However, it uses RMI for the workflow enactment service, and a web-based user interface instead of web services tools.

NetTraveler is a framework for web services collaboration, orchestration and choreography in peer-to-peer autonomic environments [10]. Its main feature is the elimination of a central coordination side running queries and autonomic query execution, embedding control information in the request and partial data results. NetTraveler seeks to integrate large amounts of

heterogeneous data and information sources. Lynx could extend NetTraveler to interact with human partners.

Ranganathan [1] proposes an architecture that integrates workflow into a pervasive computing environment. This architecture provides a system that generates a customized workflow that describes how various services should interact with one another. However it lacks a mechanism for human interaction through email. Human interaction is allowed via a custom web interface generated by a user-interface web service thus requiring significant development effort and specialized user training.

6. CONCLUSIONS AND FUTURE WORK

We have presented the design and implementation of Lynx, an open architecture that extends web service based workflow engines with human interaction via email. Lynx uses a general purpose email messaging architecture to interact with human partners by using the BPEL language for specifying business process workflow behavior based on web services. Lynx uses XForms to minimize the amount of custom code required to implement the user interfaces.

We have completed the implementation of an initial Lynx prototype and are actively pursuing the deployment of systems based on Lynx at several government agencies in Puerto Rico, including the Registry of Deeds. We are currently working on augmenting Lynx to support more documents, and providing privacy and security. More information regarding Lynx can be found at www.admg.ece.uprm.edu.

Experience testing our prototype on the Advanced Data Management Laboratory, at the University of Puerto Rico at Mayagüez, suggests that Lynx may dramatically facilitate the development and maintenance of complex interactive process oriented web applications by significantly reducing the amount of plumbing and custom GUI code that current popular approaches such as Struts or Java Faces require. A series of experiments and analyses will be conducted to assess the performance of the system in terms of throughput and system load capacity in order to verify that the architecture achieves acceptable performance for the type of applications supported. Analysis of the amount of code required for the application will be compared with estimates of that required for alternative architectures in order to test the hypothesis that the Lynx architecture can reduce the amount of custom code required for each new application.

Finally, we expect exciting results such as more agility in government transactions with the use of the outcomes of this work.

7. ACKNOWLEDGMENTS

The UPRM Digital Government project is funded by NSF Grant EIA-0306791.

8. REFERENCES

- [1] A. Ranganathan, S. McFadding. Using Workflows to Coordinate Web Services in Pervasive Computing Environments. *IEEE International Conference on Web Services*. 2004.
- [2] ActiveBPEL, <http://www.activebpel.org/>
- [3] Axis, <http://ws.apache.org/axis>
- [4] Apache Tomcat. The Apache Jakarta Project. <http://jakarta.apache.org/tomcat/>
- [5] Apache Xindice. <http://xml.apache.org/xindice/>
- [6] Business Process Execution Language for Web Services Version 1.1, BEA, IBM and Microsoft, May 2003, <http://www-106.ibm.com/developerworks/library/ws-bpel/>
- [7] Chiba. <http://chiba.sourceforge.net/>
- [8] D. Chakraborty, H. Lei. Pervasive Enablement of Business Processes. *Proceedings of the Second IEEE Annual Conference on Pervasive Computing and Communications*. 2004.
- [9] D. Ganesarajah, E. Lupu. Workflow-based composition of web-services: a business model or a programming paradigm? *International Enterprise Distributed Object Computing Conference*. 2002.
- [10] H. Caituiro-Monge, M. Rodríguez-Martínez. Net Traveler: A Framework for Autonomic Web Services Collaboration, Orchestration and Choreography in E-Government Information Systems. *Proceedings of the IEEE International Conference on Web Services*. 2004.
- [11] IBM Business Process Execution Language for Web Services Java Run Time (BPWS4J). <http://www.alphaworks.ibm.com/tech/bpws4j>
- [12] M. Dubinko, L. Klotz, R. Merrick, T. V. Raman. XForms, <http://www.w3.org/TR/xforms/>
- [13] M. Dubinko, *XForms Essentials*, O'Reilly, Sebastopol, CA, 2003
- [14] S. Graham, D. Davis, S. Simeonov, et.al., *Building Web Services with Java: Making sense of XML, SOAP, WSDL, UDDI*, Sam's Publishing, Indianapolis, Indiana, 2004.
- [15] S. Weerawarana, F. Curbera, F. Leymann, T. Storey, D. F. Ferguson, *Web Services Platform Architecture*. Prentice Hall, Upper Saddle River, NJ, 2005
- [16] S. Whitaker, V. Bellotti, P. Moody, Revisiting and Reinventing Email. *Human-Computer Interaction*, Volume 20, Numbers 1 and 2, 2005.
- [17] T. Podgayetskaya, W. Stucky. A Model of Business Process Support System for E-Government. *International Workshop on Database and Expert Systems Applications*. 2004.
- [18] Web Services Activity. <http://www.w3.org/2002/ws>
- [19] XML Schema. <http://www.w3.org/XML/Schema>
- [20] XPath. <http://www.w3.org/TR/xpath>
- [21] XUpdate. <http://xmldb-org.sourceforge.net/xupdate/>
- [22] Z. Liang, R. Wong. A Lightweight Mobile Platform for Business Services Networks. *Proceedings of the IEEE EEE05 international workshop on Business services networks*, 2005.

Argos: Dynamic Composition of Web Services for Goods Movement Analysis and Planning

José Luis Ambite, Genevieve Giuliano, Peter Gordon,
Mountu Jinwala, Dipsy Kapoor, LanLan Wang

University of Southern California

ambite@isi.edu, giuliano@usc.edu, pgordon@usc.edu,
jinwala@isi.edu, dipsy@isi.edu, lanlanwa@usc.edu

Qisheng Pan

Texas Southern University
pan_qs@tsu.edu

1. INTRODUCTION

This Project Highlight describes Year 3 activities of our Argos research. The purpose of the research is to develop a flexible data query and analysis system based on the web services paradigm. Our application domain is metropolitan goods movement. The project began in August 2003. We seek to blend computer science and social science approaches by developing new data integration tools and applying them to social science research problems. The research has three objectives: 1) to advance computer science research by developing an expressive web services description language and techniques for dynamically composing web services, 2) to develop and conduct test applications of an intra-metropolitan goods movement flow model using web services in cooperation with government partners, and 3) to use the model to conduct social science research on intra-metropolitan economic linkages and spatial structure. The approach to web service composition is general and can be applied to other scientific data gathering and analysis tasks.

The Computer Science Research Problem Significant advances have been made in both data integration and web services, but limitations still exist. Query evaluation systems in data integration systems are composed of classical relational algebra operations, but do not include arbitrary computations as required in scientific workflows. Current web services tools require manual composition. Our approach is to combine the benefits of data integration and web services by developing an architecture for automated web service composition.

The Domain Research Problem Goods movement is becoming an increasingly important phenomenon in metropolitan areas, yet our ability to obtain information, model, and plan for urban freight flows is limited. Many limitations are related to data: 1) firm specific information is proprietary, 2) available data sources are incomplete and gathered in different units or time periods or geographic scale, 3) detailed information on flows on the highway network are costly to obtain and therefore scarce, 4) traditional methods for modeling freight flows depend on highly disaggregate data which can only be obtained via field survey. Our approach is to use data from widely available secondary sources to develop goods movement monitoring and modeling tools. This requires efficient computational tools that can integrate data from heterogeneous sources and process it to generate new results.

2. YEAR 3 ACTIVITIES

Over the past year we have focused on three tasks: 1) using new data sources and refining our goods movement workflow and model; 2) developing a method for automatic composition of

computational work flows and implementing the entire goods movement workflow; 3) developing accessibility measures based on workflow intermediate products and obtaining property transaction data for testing accessibility measure

2.1 New Data Sources

In previous work we developed a model for estimating intra-regional freight flows on the highway network [5]. Our objective was to develop a model that was behaviorally robust, transferable across metropolitan areas, and used data from widely available sources. Tests of the model gave encouraging results [1]. However, the model included many assumptions due to lack of data, and relied on some unique data sets (one-time surveys). Models relying on such sources are not easily updated or transferable. We have developed a freight flow model based almost entirely on commonly available, inexpensive secondary data sources. The major research steps are: 1) utilize a regional input/output transactions table to estimate zone level intra-regional commodity-specific trip attractions and trip productions; 2) estimate commodity-specific interregional and international trip attractions and productions for zones with import/export nodes (airports, seaports, etc.); 3) create a regional origin-destination matrix using estimates from the previous two steps; and 4) load the O-D matrix on a regional highway network with known passenger flows.

Our case study is the Los Angeles CMSA; it includes the ports of Los Angeles and Long Beach, together the largest container port in the US, as well as the nation's second largest air cargo hub. Consequently inter-regional and international commodity flows represent a large portion of goods movement within the region. Last year our work focused on the estimation of inter-regional and international trip attractions and productions (supplies and demands for commodities).

Using various data sources required a method for reconciling different product classification systems. We developed conversion matrices for the classifications used in our key data sources. We implemented a generic product conversion operation in the planner that given a data relation expressed in one product classification and a translation table from such product classification to another as inputs, produces a data relation expressed in terms of the second product classification as output. The planner automatically inserts instances of this operation whenever needed.

Producing the O-D matrix requires combining data to estimate all productions that are consumed in the region, outside the region but inside the US, and outside the US. Similarly, we must estimate all goods consumed in the region that are produced in the region, outside the region but inside the US, and outside the US.

These productions and attractions must be located within the region to the zone level, must be expressed in tonnage units, and must be allocated to specific modes (truck, rail, air water). We have generated an updated O-D matrix and conducted the highway network simulation. Preliminary results are promising.

2.2 Automatic Workflow Composition

A major goal of this research is to develop methods for automatically compose computational workflows. In this year we achieved four important milestones. First, we have developed a expressive modeling methodology to describe the contents and sources and the inputs and outputs of data processing operations. These descriptions are expressed in a first-order logic language, using object-oriented modeling principles in the style of description logics. Second, we have developed an ontology for the commodity flow domain that covers all the data products and operations that we have encountered and that is easily extensible. Third, we have designed and implemented a planning approach that fully exploits these descriptions. Finally, we have implemented an execution system that evaluates the automatically generated plans.

Our modeling methodology uses a domain ontology to describe the data products in sources and operations. We represent these data products as relations that have a formal definition, a logical formula composed of concepts (classes) from our domain ontology. A novel feature of our approach is that we associate each data product relation with one or several concept definitions to speed up subsumption reasoning. Our ontology follows the principles in our previous work [2], but the definitions are significantly more precise. We have expanded the scope and the number of operation modeled. We use the very expressive PowerLoom knowledge representation and reasoning system [3,4].

Our planning approach is a regression search from the user request using the available operators until reaching appropriate sources. The main reasoning step is satisfying the input of an operation with output of another operation or source. The system proves that such input and output are equivalent using subsumption. The Powerloom logic, though incomplete, can prove subsumption quite efficiently. Currently, we can generate plans of about 50 operations in under a minute.

Finally, the system translates the automatically generated plans into a form suitable for execution. Originally, we implemented our execution system in BPEL. However, since our sources and intermediate results are quite large (processing about 2 million tuples is common), we implemented many of the operations on top of a database system (MySQL) for efficiency. We plan to explore efficient execution of web services for large data sets, possibly moving to a grid services infrastructure.

2.3 Accessibility Measures

The third part of the Argos research project is to use some of the intermediate products of the computational workflow to compute accessibility measures based on economic activity. The workflow allows us to generate access measures based on employment, commodity flow, or economic attractiveness (trip ends), all by industry sector. Development of such detailed measures at a fine geography has not previously been possible. We want to examine

the relationship between these access measures and urban spatial structure. We obtained property transaction data for 2000 for this purpose. We have developed a variety of access measures, including some based on employment centers (part of a related research project). Estimation of models testing the significance of various access measures is in progress.

3. Conclusions

We have now completed about 2/3 of this research project. We have accomplished much of what we had intended, and the collaboration between computer science and social science colleagues has been both challenging and productive. Identification of suitable and compatible secondary data sources required more time and effort than anticipated. Changes in data sources required changes in the ontology and workflow, which delayed the computer science work. After gaining greater knowledge of the domain, our initial representation and reasoning system, Triple [6], proved insufficient. So, we turned to a more expressive logic: Powerloom, which is significantly more expressive and adaptable [3, 4]. We experienced further delays in obtaining property transaction data; however we are now making good progress on this part of the analysis.

We have begun to generate research contributions: a demo at dg.o 2004; paper at dg.o 2005 [2]; poster presentation on secondary data sources at the Commodity Flow Survey Conference (2005); journal article on freight model structure [5] We have submitted a paper and a poster demo on different aspects of the research to dg.o 2006. We plan to submit research papers to AAAI and the ISWC conferences on our automatic composition approach.

REFERENCES

- [1] J. L. Ambite, G. Giuliano, P. Gordon, Q. Pan, S. Bhattacharjee, (2002) Integrating heterogeneous sources for better freight flow analysis and planning, *Proceedings of the 2nd Annual National Conference on Digital Government Research (dg.o 2002)*, Redondo Beach, CA, 2002..
- [2] J. L. Ambite and M. Weathers. Automatic composition of aggregation workflows for transportation modeling. In *Proceedings of the 6th Annual National Conference on Digital Government Research (dg.o 2005)*, Atlanta, Georgia, 2005.
- [3] H. Chalupsky and T. Russ. WhyNot: Debugging failed queries in large knowledge bases. In Proceedings of the Fourteenth Innovative Applications of Artificial Intelligence Conference (IAAI-02), pages 870–877, Menlo Park, 2002. AAAI Press.
- [4] R. MacGregor. A description classifier for the predicate calculus. In Proceedings of the Twelfth National Conference on Artificial Intelligence, Seattle, WA, 1994.
- [5] Q. Pan, Freight data assembly and modeling, *Transportation Planning and Technology*, 29(1), forthcoming, 2006.
- [6] M. Sintek and S. Decker. TRIPLE: A query, inference, and transformation language for the semantic web. In *International Semantic Web Conference (ISWC)*, Sardinia, June 2002.

Data Processing Workflows in the Social Sciences: Representation and Automatic Generation (Abstract)

José Luis Ambite
Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292
ambite@isi.edu

Dipsy Kapoor
Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292
dipsy@isi.edu

Mountu Jinwala
Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292
jinwala@isi.edu

1. INTRODUCTION

Much of the work of social scientists and government practitioners is consumed by accessing, collating, and analyzing data. This is particularly true in the planning and economic modeling agencies. Unfortunately, there is a severe lack of tools to facilitate this process and much of the integration is done manually by ad-hoc methods. Moreover, raw data are of limited utility. Usually these data are the input to models of more complex phenomena that produce additional data of interest. For example, in our commodity flow domain, we derive truck traffic along specific highway links within a metropolitan area, based on quite far-removed raw (source) data such as employment, imports into and exports out of the region, etc, by using a complex workflow of operations.

The goal of the Argos project is to improve this state of affairs by providing a framework to (1) describe data products and data processing operations so that they can be shared and reused and (2) automatically generate new data products on demand by automatically composing data processing workflows using available sources and operations. In this abstract, we outline our solutions to these two challenges. A novel contribution is that our approach can automatically insert operations that make the inputs and outputs of different operations compatible (so called *shims* [1]).

2. DOMAIN REPRESENTATION

One of the major challenges to automating computational workflows is understanding the data products present in available sources and operations. After some effort, a domain expert can accurately describe such data. However, such descriptions are rarely recorded and much less formalized unambiguously. To ensure a clear data semantics, we have developed a formal ontology for our goods movement domain.

The Argos Ontology represents the concepts and relations of our transportation domain in the expressive Powerloom language, a first-order logic with recursion [3]. Powerloom provides logical inference services to the Argos system, in particular, *subsumption*. Subsumption proves that membership in a class (or relation) logically implies membership in another class (or relation). First-order logical inference is undecidable, hence Powerloom is incomplete. Nevertheless, Powerloom is specially optimized to compute subsumption and Powerloom proves the inferences required by our (quite expressive) ontology efficiently.

Figure 1 shows some sample concept, instance and rule definitions of the Argos ontology. The Flow concept rep-

resents a transfer of a product between two geospatial areas (from an origin to a destination) using a transportation mode measured during a particular time interval. For example, an instance of Flow is the domestic exports by air (a *TransportationMode*) of Pharmaceutical and Chemical products from the Los Angeles-Riverside-Orange County, CA, Consolidated Metropolitan Statistical Area (LACMSA) in 2000, which amounts to 2226 million US dollars.

The ontology also encodes information about well-known entities in the domain. For example, Figure 1 shows the fact that Los Angeles County (g-LA) is geographically contained in (is a *geoPartOf*) the LACMSA area, as well as Ventura County (g-VT), something not immediately apparent from the LACMSA name. Finally, the ontology includes rules clarifying the semantics of the concepts and relations. For example, the *recursive* rule in Figure 1 specifies the transitivity of geospatial containment (*geoPartOf*).

```
(DEFCONCEPT Flow (?x) :<=> ;; concept definition
  (exists (?o ?d ?p ?t ?u ?m ?v)
    (AND (Data ?x)
      (hasOrigin ?x ?o) (Geo ?o)
      (hasDestination ?x ?d) (Geo ?d)
      (hasProduct ?x ?p) (Product ?p)
      (hasTimeInterval ?x ?t) (TimeInterval ?t)
      (hasUnit ?x ?u) (Unit ?u)
      (hasMode ?x ?m) (TransportationMode ?m)
      (hasValue ?x ?v) (Number ?v) )))

(USGeo g-LACMSA) (USCounty g-LA) ;; instance assertions
(geoPartOf g-LA g-LACMSA) (geoPartOf g-VT g-LACMSA)

(forall (?x ?z)
  (=> (exists (?y)
    (and (geoPartof ?x ?y) (geoPartof ?y ?z)))
    (geoPartof ?x ?z))))
```

Figure 1: Argos Ontology: Sample Definitions

Using this formal ontology we describe the data products provided by sources, and required or computed by data processing operations. For example, Figure 2 shows the description of the contents of a table that provides the number of jobs in 2000 for each Traffic Analysis Zone (TAZ) contained in the LACMSA, for products categorized following the 1999 Standard Industrial Classification (SIC) codes (with a granularity of 4 digits). Since the data description uses the logical biconditional (\leftrightarrow), it means that the table has the *complete* set of tuples that satisfy the relation definition, i.e., the table contains values for *all* the products of type *Product-sic-4-1999* for *all* the TAZs in the LACMSA.

```

(defrelation Data-Rel-Employment-2000-LACMSA-TAZ-SIC
  ((?county USCounty) (?jobs Number)
   (?p Product-sic-4-1999)(?taz TAZ)) :<=>
  (exists (?o)
    (AND (Measurement ?o)
      (hasProduct ?o ?p)
      (hasGeo ?o ?taz) (geoPartof ?taz ?county)
      (geoPartof ?county g-LACMSA)
      (hasUnit ?o u-NumberOfJobs)
      (hasTimeInterval ?o 2000)
      (hasValue ?o ?jobs))))

```

Figure 2: Sample Data Product Description

3. AUTOMATICALLY COMPOSING DATA PROCESSING WORKFLOWS

We assume that sources and operations have been developed independently. For example, the operations may be web services and the sources external databases. This presents two challenges. First, each source may use a different schema. Second, the data produced by a source or operation may not be input directly into other operations, but need some kind of transformation.

Argos addresses both these challenges. First, it resolves the semantic heterogeneity by mapping all data products to a common ontology. Second, Argos provides a library of domain-independent operations and a framework to define generic domain-dependent operations that can bridge the differences between the input required by some operation and the output provided by another (*shims* [1]).

3.1 Operations

A data processing operation is represented by its input/output signature. Each input or output is described by a relation definition in the ontology (e.g., Figure 2). An operation can have multiple inputs and outputs. There are three types of operations supported in Argos:

Sources and Domain-Dependent Operations. The inputs and outputs are described by data relations predefined in the ontology. A source is a domain-dependent operation that requires no inputs.

Domain-Independent Operations. In order to bridge the inputs and outputs of different operations, Argos provides a set of built-in domain-independent operations similar to the relation algebra operations: selection, projection, join, and union. However, the system uses the background ontology to prove that inserting such operation is semantically valid. We illustrate the process with selection. The other operators are analogous.

Assume that the planner wants to obtain the employment data for the TAZs in Los Angeles County, but the available operators can only produce the employment data for all the TAZs of the LACMSA (cf. Figure 2). Using the ontology, the system reasons that since Los Angeles County is geographically contained in the LACMSA (cf. Figure 1) the desired relation is a subset of the available one. After also checking that county is an output attribute of the provider relation, the system will insert a selection operator.

Generic Domain-Dependent Operations. There are a variety of operators that lie between the completely domain-specific operators (described by predefined datasets), and the (relational-algebra-like) domain-independent operators. Product conversion is a prime example of a generic, but domain-dependent operator.

Economic data is reported in a variety of product/industry classifications (NAICS, SCTG, SIC, ...). Our project social scientists have created translation tables between several of these classifications. Thus, we added a generic product conversion operator to the Argos library. To satisfy a data request for products in a classification C2, this operator subgoals on obtaining the data in classification C1 and a translation table from C1 to C2.

3.2 Planning

The Argos planner automatically generates a workflow of sources and operations in response to an user data request. Our planner performs a regression search in plan space (cf. [2]), starting with the user data request as goal, until it finds a plan that computes such request using available operators where all data inputs to operators can be satisfied by other operators or by the available sources.

The basic plan refinement step is to satisfy an operator input with the output of another operator or source. In order to ensure that the input and the output data relations are semantically compatible, the planner performs an equivalence test, that is, it checks subsumption in both directions.

4. DISCUSSION

The Argos planner and executor are fully implemented. We have developed a core ontology for our transportation domain with about 40 concepts and 10 relations. We also described 17 sources and 11 domain-specific operations.

Our initial experiments are promising. The planner generates a workflow with 17 operations (7 sources, 4 domain-specific operations and 6 product conversions generated on-the-fly) in about 20 seconds. A larger workflow with 54 operations (17 sources, 11 domain-specific operators, 8 product conversions, and 18 projection operators) takes 2 minutes and 44 seconds.

In the immediate future, we plan to scale the number of operators and sources in our commodity flow domain. In addition, we want to broaden the scope of work to other questions of spatial urban structure.

5. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Award No. EIA-0306905. We would like to thank the social scientists of the Argos project for their contributions: Genevieve Giuliano, Peter Gordon, Qisheng Pan, LanLan Wang, and JiYoung Park; as well as our government partner agencies.

6. REFERENCES

- [1] D. Hull, R. Stevens, P. Lord, C. Wroe, and C. Goble. Treating shismatic web syndrome with ontologies. In *Procs. 1st AKT workshop on Semantic Web Services (AKT-SWS04)*, Milton Keynes, UK, 2004.
- [2] C. A. Knoblock. Building a planner for information gathering: A report from the trenches. In *Procs. 3rd International Conference on Artificial Intelligence Planning Systems*, Edinburgh, Scotland, 1996.
- [3] R. MacGregor. A description classifier for the predicate calculus. In *Procs. 12th National Conference on Artificial Intelligence*, Seattle, WA, 1994.

SESSION 8B

FEDERAL AGENCIES AND THE WEB

Moderator

J. Ramon Gil-Garcia, University at Albany/SUNY, USA

Titles and Authors

Federal Agencies and the Evolution of Web Governance
Mahler, Julianne; Regan, Priscilla M.

Data Confidentiality, Data Quality and Data Integration for Federal Databases
Karr, Alan F.

*Integrating Data and Interfaces to Enhance Understanding of Government Statistics:
Toward the National Statistical Knowledge Network Project Briefing*
Marchionini, Gary; Haas, Stephanie; Plaisant, Catherine; Shneiderman, Ben

Federal Agencies and the Evolution of Web Governance

Julianne Mahler

Department of Public and International Affairs
George Mason University
MSN 3F4
Fairfax, VA 22031
jmahler@gmu.edu
1 703 993 1414

Priscilla M. Regan*

Department of Public and International Affairs, GMU
Fairfax, VA 22031
Science & Society Program,
National Science Foundation
pregan@gmu.edu
1 703 993 1419

Abstract: Over the last ten years federal agencies have undergone a fundamental transformation in the way they manage programs and internal administration, relations with Congress, and dealings with clients and citizens. Agencies now operate in sophisticated electronic environments of email, electronic documents and filings, Intranets, and Internets. This paper describes and analyzes the emergence of “web governance.” Briefly, web governance is concerned with the control of content and design for agency websites. Here we focus on the processes by which web governance decisions are being made by individual federal agencies. We discuss preliminary information about how agencies are responding to Office of Management and Budget’s new web content rules.

General Terms: management

Keywords: digital government, website control, governance

1. INTRODUCTION**

The transformation to e-government has recently generated calls within federal agencies and across the federal government for tighter control over the content and management of federal agency websites, more uniform standards for presentation and accessibility, and more centralized procedures for tracking and approving material on the websites. These new policy efforts include the creation of more elaborate procedures for approving website content in many federal agencies. Since 2002, the Office of Management and Budget (OMB) has also been charged with overseeing the creation and implementation of government-wide standards to guide the look and performance of federal websites, and more specific procedures for tracking and approving the content of sites are being created at the agency level.

In this paper we begin to analyze the development of these new governance efforts. How are decisions about the control of web content decisions being made within agencies? What kinds of processes and influences shaped the emerging guidelines and approval procedures? We seek to describe the emergence of what is now being termed “web governance,” the authoritative control of content and design for agency websites. At stake may be a significant change in how decisions about website content are being made and who makes them.

The term “web governance” encompasses a broader and more theoretically grounded orientation than terms such as information resource management, knowledge management, or technology management. “Web” includes not just the technology and content, but also providers, users, and networks. Similarly “governance” embodies the evolution and development of rules, processes and structures within agencies, across agencies, and most centrally, government-wide. These concerns are more centrally rooted in traditional public administration literature than they are in information policy or technology management. Fountain has termed this larger enterprise the building of the “virtual state,” a structural and institutional approach focusing on the interdependence between organizations – including rules, procedures and actors – and technical systems [1]. In this paper we begin to investigate what decision-making and problem-solving processes are emerging, which governmental actors are involved and the roles they play, how these processes are organized, and what types of decisions are being made. In this respect, “web governance” is closely aligned with Fountain’s term “networked governance” [2].

2. THE ORGANIZATIONAL SETTING OF WEBSITE GOVERNANCE

What should we expect the development process for website governance tools to look like? Since the 1960s a shift from hierarchical and mechanistic to organic and decentralized organizational processes has been associated with high technology. Current research on the kinds of collaborative work settings and problem solving capacities of internets and intranets shows how internets and intranets fulfill some of these early ideas about lateral communication and knowledge management [3]. Self-organizing, learning organizations are thought to be fostered by the open, wide dissemination of information made possible with internets and intranets. Research on the evolution of creative, interactive intranets sees them as emerging from the bottom up, in several places simultaneously, rather than in an orderly, centralized process [4].

Peristeras, Tsekos and Konstantinos note that with the advent of widespread information and communication technologies that made e-government possible, public administration is approaching a paradigm shift [5]. Rather than relying on rule-governed hierarchical control to coordinate effort, work and workers become increasingly self-organizing, using the networks of relations made possible by new, widely distributed electronic connections [6]. Agency actors may take on roles of innovators, boundary spanners, risk takers and entrepreneurs [7]. Networks replace hierarchies as workers find ways to partner with agencies and other private and non-profit partners to achieve the public purpose [8]. Large scale routine technologies are replaced by communities of practice, in which knowledge of how to achieve the desired ends is located in tacit practice, shared by those connected by interests and perhaps only secondarily by formal organizational settings [9]. These communities are based on the shared understanding of how to do work based on expectations built up over past interactions [10].

Ideas about self-organizing systems rest on several sources including complexity theory [11] and ideas about the coordinating properties of markets [12]. Self-organization offers an alternative to hierarchy and strategic choice as the force behind organizational innovation and change. Rather than reacting to negative feedback, the organization's "futures emerge unpredictably from the interactions between agents in conditions of nonequilibrium and disorder" [13].

In this view, the explosion of pages of rich and innovative content on federal agency websites is, at least in part, the consequence of the unplanned freedom to post. In addition, we would expect that agency rules, including rules about postings, would take on a different character from the hierarchical and controlling forms seen in traditional accounts of organizations. In e-government settings, rules that constrain behavior may be less typical than rules that address problems and guide solutions [14]. The primary impetus for solution-guiding rules is related to performance enhancement rather than accountability. E-government initiatives have primarily concentrated on performance enhancement and hence the development of solutions. The solution-guiding approach enables managers to create organizational solutions to simplify their decisional processes and to resolve challenges the organization is facing. Such an approach emphasizes organizational learning, embodying the norms and best practices that have evolved most often from the bottom up.

If this dispersed and self-designing process characterizes emerging efforts to create guidance for agency website content decisions, we would expect it to operate on the basis of wide participation, collaboration rather than control, and self-organizing procedures rather than tightly held, top-down control over participation. In fact, we did see such a process in the development of OMB's 2005 guidelines for federal agency website content and appearance [15]. In drafting these guidelines, OMB created the Interagency Committee on Government Information (ICGI) and subsidiary working groups that were allowed considerable autonomy and operated in a clearly collaborative and open fashion. The result was a process that was largely "owned" by the agency staff who had developed expertise and interest in e-government through direct experience in their

own agencies. A basis for this development process was the volunteer-based Federal Content Managers Forum, originally a small lunchtime gathering of early content managers struggling with website issues in their agencies. This Forum now has a listserv of 400, and its membership provided the basis for identifying the content managers to serve on the interagency council. This unusual source for the ICGI, composed of agency personnel who were informally and voluntarily drawn to the problems of content management, is an example of the emergence of new self-organizing processes in public management.

Do we see the decentralized, self-organizing process at the agency level as well? Is a similar process occurring as agencies move to develop web content governance? Or is agency change still better characterized as strategic and contingent. The stakes for content management decisions and policies are high. The websites are seen as an increasingly critical agency resource for managing relations with the public and other branches of government [16]. There may be struggles over control of this resource rather than collaboration and openness.

An alternative to the self-organizing view of organizations are more traditional strategic and contingent views of program and structure decisions. Environmental imperatives [17] and resource dependency [18] drive choices about organizational policies and designs. These are strategic decisions, whose result is intended and deliberate, though not always successful. Powerful professional collations within organizations vie for control of scarce resources [19]. Organizations are institutions that act at the behest of those who own or control them. They are instruments of governments or boards of directors. In this view, website postings for governmental agencies are another tool to be used to further the agenda of the government, though there may be conflict over the control of this resource.

This research contrasts two views of how organizations create and post web content. One views organizations as very complex systems of actors whose joint efforts result in the rich resource we call the internet. As in markets or democracies, the most valued results emerge spontaneously from these undirected actions. The other broad perspective views these same organizations as instrumental entities, whose benefits are gained through strategic action and political control.

Our agency case studies give us an opportunity to look for evidence of these views in examining process by which web content approval process at the agency level. In a small way, this research tests the depth to which the new collaborative management paradigm can survive serious competition over a valued agency resource.

3. METHODOLOGY

The research conducted to examine the development of web governance procedures includes two sets of interviews with agency personnel concerned with web content decisions. First, to investigate how agencies were creating their own web governance procedures, we conducted interviews with staff responsible for content management at various offices in seven agencies as well as agency staff who participated in several formal and informal

interagency councils concerned with content management issues and web governance during the summer and fall of 2004. In addition, we perused publications and online documents setting out the developing procedures within agencies such as rules for posting decisions and memorandums of understanding about authorization to post content. It is important to note, however, that written rules have not been developed in all cases. We also gained access to interagency recommendations and government-wide guidelines.

To investigate the development of content management procedures within agencies, we selected seven agencies with varying levels of website development and sophistication as judged in West's fourth annual rankings of federal websites. These rankings are based on the quality of online publications, online databases, audio and video clips, advertisements, fees, privacy and security policies, comment forms, readability level, and presence and number of online services [20]. Although these features are not directly relevant to our interest in content management for agency websites, they do provide an overall measure of the sophistication and professionalism of a website. Using the West rankings, we selected two agencies rated most highly (Federal Communications Commission, FCC; Housing and Urban Development, HUD), three in the middle (Food and Drug Administration, FDA; Department of Transportation, DoT; Environmental Protection Agency, EPA) and two at the bottom of the rankings (Equal Employment Opportunity Commission, National Labor Relations Board). The West rankings allow us to compare agencies with well-developed web presence to those just beginning to offer more kinds of information online. Variation in stages of development may be important to decisions about content management and governance.

Interviews were conducted with selected agency staff responsible for information technology, information management, program content, and public relations in each of our agencies. We interviewed a total of 18 agency officials in 10 offices and 7 former and current staff at OMB. In some large, multi-divisional Departments, we interviewed web content staff in program divisions as well as in upper level public relations staff, sometimes located in the top administrator's offices. Subsequent phone calls and emails allowed us to clarify and elaborate on the information gleaned during the interview.

A second set of data was collected in the fall of 2005 to investigate agency reactions so far to the OMB Guidelines. We conducted an informal email survey of the members of the Federal Web Content Managers Forum, described below. Eleven representatives from seven cabinet departments and two smaller independent agencies responded to our online request for information. The request was sent via email in September 2005 to the members of the Forum's listserv. We asked about twelve issues including:

- Respondents' reports of how the OMB guidelines were viewed by others in their agency, and whether they thought these views were widespread in the agency;
- The progress made to date in implementing the guidelines and the identity of the unit that was taking the lead in implementation.
- The degree of controversy encountered in the implementation effort and whether the process was best described as collaborative or centrally controlled

- What online resources the agency had consulted to help implement the guidelines, including the OMB website, webcontent.gov, the best practices pages in other agencies, etc
- whether the agency had its own formal or informal web content group and whether the group has created its own web content policies.

Respondents were told that their agency and their names would not be revealed. For purposes of analysis, we indicate where the cabinet level agencies generally fall in the 2005 West ratings of the top 60 federal agency websites: three agencies are in the highest tier of West ratings (above 70); four are in the medium 50-70; and two in the lowest tier [21]. The independent agencies were not ranked in West's 2005 analysis. The respondents to these questions were not the same as the ones we interviewed a year earlier.

4. AGENCY GOVERNANCE INITIATIVES

At the agency level we investigated where the decisions about web content are being made and what policies are appearing to govern who can create and post content. We questioned actors about the approval processes emerging for web content and what kinds of standards for content were emerging. Together these issues constitute what we mean by governance, the authoritative control of content and design for agency websites. We found that the approval processes and the principles that actors thought were served by the approval process vary widely among the agencies in the study. Two issues that capture much of what we found are the degree of centralization and the location of web content authorization.

The results from the 2004 agency interviews showed there were distinct differences in perceptions about where web content approval decisions were made [22]. In almost all cases top-level agency web content managers reported that the program offices had almost complete autonomy over the initiation and control of the content of the sites. However, virtually without exception the web posters or webmasters in *subordinate* program offices identified an elaborate and careful hierarchical procedure for review and approval of content. Some upper level managers are trying to prevent website posting from becoming (or remaining) a "free for all," as one respondent put it, and some program officers are trying to exercise their accustomed autonomy in these decisions.

A number of respondents described tension or conflict in their agency over the control of the website content. Behind these struggles are disagreements over who the agency's stakeholders are, how they can be served, and how the agency's mission is advanced through the use of the public website. Interviews revealed debates between those who thought that the website should be used for public relations and those who saw it as a tool for communicating program information to the public.

Conflicts also emerged over the location of the web content manager within the agency. Increasingly the content functions described above are now located in Public Affairs or the Director's office rather than with IT or IRM functions, reflecting a shift away from viewing the Website as a technical feature of the

agency. In two cases we heard how the IT division gave up or did not seek to control web content decisions only to realize later what a significant organizational resource such control was. In each case efforts by IT staff to wrest control from content managers were unsuccessful. Other conflicts arose between program officials in the subdivisions of the agency and public affairs offices at the agency director's level.

These conflicts and the typical uncertainties about who should authorize important postings led some of the agencies in our study to create formal procedures or bodies for making posting decisions. In others, as noted, the program actors seem to have greater independence, leading to a "wild west" approach in the words of one respondent. This freedom is said to create inconsistent policy and legal positions. Some agencies have confronted these problems directly by creating governance structures to rein-in "renegade" sites, as officials in one agency respondent put it. Three have developed procedures or institutions to govern content decisions, while three more are at varying stages of trying to create such structures.

In one large regulatory agency, a governance committee with formal authority to reorganize the website has been created. The entire Web Working Group consists of about 500 staff and contractors, but a new formal governance structure is emerging with two representatives each from public affairs, the public information office, the web steering committee, program office representatives and regional representatives. The group is drafting website design principles and a management structure to improve the integration of information across the presently segmented website. Program participants say it will enhance the usefulness of the site for citizens, and will help ensure a consistent message across the site.

But in another large regulatory agency, each subdivision has its own webmaster. An internet policy group emerged in 1996, one year after the emergence of the website itself, to share information and help make the site more user-friendly. Originally, the group was more focused on technical than design issues. Despite its name the group is advisory, not policy making. In 2001 the work was focused on providing a common look and format, and greater usability for the agency's audience of technical users. The group continues in an advisory capacity.

At a third large regulatory agency, an Internet Work Group is directed from the public affairs staff in the Commissioner's office. It has created a clear set of procedures for creating web pages that meet the agency's design standards and are well coordinated with other agency materials. The policy guidelines are explicit about not imposing onerous clearances on projects. They state "We do not want to place unnecessary constraints on the design and management of the Web sites operated by [the agency]." The guidelines themselves describe standards for formatting and coordinating web information. They describe the clearance process through the Office of Public Affairs, and they offer guidance on how to prepare program clearance forms.

In three other agencies, governance efforts are just beginning. A small agency tried four or five years ago to set up a task force to create a more useable site that would educate citizens in how to use the services the agency provides, but lack of support

from top-level management blocked the effort. Our respondent characterized the website currently as static. Another small agency has a content management committee composed of top program management and presidential appointees. A larger regulatory agency had a web council of content managers with advisory status. They pushed for the recent hiring of a new content manager and helped establish a policy making group. This group includes officially designated web liaisons from the subdivisions of the agency and an executive committee with greater authority to make web policy. In other offices, procedures and clearances are managed but no power-sharing arrangements appear to have emerged so far.

5. EARLY AGENCY REACTIONS TO OMB GUIDELINES.

In addition to agency-level web content governance efforts, OMB has issued government wide website guidelines. The current role for OMB's oversight of the development of federal websites evolved over the past 25 years from a series of legislative measures and executive orders beginning with the Paperwork Reduction Act (PRA) of 1980. In 1995 OMB acquired greater responsibilities for developing federal information technology (IT) performance standards and procurement through use of its budget oversight capacities. The Clinger-Cohen Act, also known as the Information Technology Management Reform Act, almost immediately amended the 1995 Paperwork Reduction Act and created a Chief Information Officer (CIO) for each agency. President Clinton's Executive Order 13011 established the CIO Council, chaired by OMB, to provide an interagency forum for CIOs to exchange information and to make recommendations about the development of effective IT management strategies. Congress went further in the E-Government Act of 2002 to establish the Interagency Committee on Government Information reporting to the CIO Council and OMB, to "to improve the methods by which Government information, including information on the Internet, is organized, preserved, and made accessible to the public." This group, mentioned earlier, permitted the guideline process to proceed in a decentralized fashion., (Information and documents on the Interagency Committee on Government Information are available online at: <http://www.cio.gov/documents/ICGI.html>) A subgroup of the Committee, the Web Content Standards Working Group, issued its recommended policies and guidelines for Federal public websites on June 9, 2004. They also developed and posted on FirstGov.gov "The Federal Web Content Managers Toolkit." This contains a compilation of laws and regulations, a list of common practices, and tools including a listserv, discussion forum and library.

How have agencies responded to these government-wide governance standards? In 2004, the actors we spoke with in agencies with larger and more fully developed websites were not concerned about the guidelines on web content developed by the Interagency Committee on Government Information and released by OMB [23]. There appear to be several reasons for this. The larger sites are already meeting and exceeding the guidelines, which in any case their web content representatives likely had a hand in producing. Of the agencies we examined, HUD, FDA and EPA in particular were well represented on the Committee.

Several noted that they only needed to tweak current practices to meet the guidelines.

Table 1 . “How do you think people in your agency view these guidelines?”

Agency View of Guidelines	Number of Responses
Useful/helpful	10
Compatible with existing procedures	6
Micromanaging	3
Burdensome	3
Offers new ideas	3
Irrelevant	2
Redundant	1
Other	2 unrealistic timeframe; solid framework in one place

A year later, eleven representatives from seven cabinet departments and two smaller independent agencies responded to our informal, online request for information about the OMB Guidelines. Based on these very preliminary findings, a few observations emerge. First, almost all of the respondents described the guidelines as “useful and helpful.” (See Table 1 for the overall responses.) However, those who agencies whose sites were ranked the highest in the West ratings all said they found the guidelines to be compatible with their existing procedures. Those from the mid-level West category, however, were much more mixed in their responses; only half of them thought the guidelines were compatible, and these respondents were also among the few, along with those in the lowest categories, to indicate that the guidelines were irrelevant, micromanaging, or burdensome. Only one of those in the lowest categories said the guidelines were compatible with existing procedures. These views were generally said to be widely held within the agency. So among these respondents, not surprisingly, there does appear to be a pattern such that the agencies with more advanced sites found the guidelines compatible with what they were already doing. This confirms what our respondents a year earlier said they expected.

Progress toward implementation has been very mixed according to these respondents. Those who said they were finished or nearly finished were all from agencies with high or mid-level scores on the West ratings. Those with the most highly rated websites all said they were finished or more than halfway, while those with the least developed sites were said they were less than halfway or halfway. When we asked how contentious the implementation process had been, most said it had not been at all controversial or only slightly controversial. Interestingly, the two respondents who said it was “too soon to say,” were from agencies with the higher West ratings. The process was also described as collaborative by 8 of the 11 respondents, including all the highest rated agencies. But five respondents, including two of the high rated agencies who said that the process had been collaborative, also described the process as directed by a central staff. Clearly the implementation route is more complex than we are able to detect with this survey.

We also tried to discover the mechanisms in place in the agencies to work on implementing the guidelines. The lead unit for six of the 11 cases was the IT office. Five respondents, all with high or middle level ratings, identified the Public Affairs Office as the leader. Recalling that there was some concern about the role of the CIO Council in designing the guidelines, this suggests that if guidelines for the control of content become more ambitious, conflict over the control of these decisions may emerge. Of those with the highest West ratings all responded that they had a formal web content group in the agency and that it had created its own policies for a web content standards and approval process. Of those in the middle group, half had a formal group with its own policies, and half had informal groups. Those in the lowest categories tended to have an informal group or no group at all. As expected, more formal agency level attention to web content issues prior to the content paved the way for smoother implementation. As was discussed above, the active agency staff involved in web content were often recruited into the ICGI working groups and the “best practices” of these agencies were often reflected in the final OMB guidelines.

Agencies were fairly uniform in the sources they consulted in implementing the guidelines. (See Table 2) Regardless of their West ratings, agencies went to the same online sources. We would have expected that the higher ranked agencies would have attended meetings of the ICGI or CIO Council. However, of the 6 agencies that reported that they had attended only 1 was rated high, 3 were in the middle and 1 was in the lowest tier (1 agency was not rated).

Table 2. “Which have you consulted in implementing the guidelines?”

Consulted	Agencies reporting
OMB website, webcontent.gov	11
Best practices on the OMB website	9
Best practices on other agency websites	8
Plainlanguage.gov website	5
Meetings of the ICGI or CIO Council	6

Respondents were also asked open-ended questions about which of the OMB Policies for Federal Public Websites were easiest to implement and which were the hardest. Eight responded to these questions. Not surprisingly respondents from all levels of the West ratings found the policies already adopted by their agencies were the easiest ones to implement. These were described by one respondent as “pretty standard legal requirements” and by another as ones “based in common sense.” In general these included the policies for protecting privacy, maintaining accessibility, using approved domains, and displaying appropriate logos. Respondents had more to say about which web content policies were hardest to implement and similar comments were made by respondents across the West ratings. Several respondents found the same policies hard to implement. One policy, mentioned by respondents from high and low ranked agencies, involved the establishment of information dissemination product schedules. This was difficult in one agency because of the lack of “an overall strategic plan” and in the other because

there was “not a distinct method for compliance.” Respondents from two medium ranked agencies considered a related policy, web records management, as difficult because it was a “diffuse and complex task” that would “require extensive effort and resources to address adequately.” The requirement to link was considered hard to implement by both low ranking agencies. Finally quality control was mentioned as hard by respondents from medium and low ranked agencies because there was not a “commitment to quality control” and because it involves “an ongoing editorial process.”

6. CONCLUSIONS

Our very preliminary findings about the government-wide guidelines suggest that there is now little resistance to the OMB rules for content. Agencies with more advanced sites view the rules as easy to comply with since they have already adopted many of the standards themselves and were likely represented in the Forum that created them. Agencies with less well developed sites, however, even though they also generally see the rules as useful, are also somewhat more likely to see them as burdensome or as an instance of micromanagement. This pattern may not be a good predictor of the future agency responses, however, among all the agencies. If the rules become more ambitious and raise tougher issues of the locus of control for sites or the process for vetting new material, they will become entangled in the same controversies that have emerged in even the most advanced agencies.

Within agencies, however, the process has been less collaborative and seems to reflect internal controversies over the mission and how the web technologies will serve that mission. Belatedly in some cases, IT offices have awakened to the value of authority to manage content. As agency websites become the most public agency presence for most citizens, the approval process for content appears to be increasingly contested. Officials in Directors’ offices, offices of public affairs, IRM offices, and, of course, program offices are vying for control. Agency level web content guidelines are emerging but apparently in a less collaborative mode than at the interagency advisory level. The concern within agencies, of course, is that the open and experimenting model of the internet posting process that characterized the early days of federal websites may be compromised and reshaped. OMB rules so far have been generally seen a benign influence, but this may not last. Since the most advanced actors were, in fact, part of the process, it is not yet clear whether the rules are leading or following agency practice.

Overall, we expected that the process of creating web content governance procedures would generally reflect the common characterization of the internet environment itself as decentralized, collaborative and self-designing. We were largely wrong. We did see elements of this kind of process in the way that OMB groups developed the guidelines as noted earlier [24]. This process created rules that appear to be only marginally intrusive on agency web content decisions. There seem to be few real objections raised to the OMB guidelines since they make only modest demands about format and virtually none, yet, about content. We did not typically see this kind of response at the agency level, however. Within the agencies, the process appears

to reflect business as usual within large bureaucratic organizations with subdivisions and levels jockeying for control.

The implications of these findings for the two models of the agency website creation process are difficult to tease out. The government-wide rules, which are not seen as burdensome by most agencies, were created through an open, bottom-up process that engaged the users of the rules in their creation. The design of these guidelines seems to reflect the self-organizing view that collaboration and a kind of democratic, contributive process will characterize the content decisions. In contrast, the instrumental view of organizations leads us to expect strategic battles for control of valued website resources, and we do largely see this at the agency level. What accounts for these differences? Two hypotheses suggest themselves. Perhaps the stakes for the actors creating the guidelines in the government-wide setting were not as high as they were in the agency context, and so a struggle for resources was not activated. If so, as the uses of agency websites are more appreciated within agencies and by government-wide authorities such as OMB or the White House, we might expect to see more onerous and constraining rules. Alternatively, agency actors may not yet see the link between a decentralized, collaborative website creation process and the innovative web content within the agency. If this is so, there is more promise for solution-guiding rules that foster collaborative governance. Greater awareness of the benefits of democratic, self government for web content may preserve this valued resource.

*This material is based upon work conducted while serving at the National Science Foundation.

** An earlier version of parts of this paper appear in Julianne Mahler and Priscilla M. Regan, “The Evolution of Web Governance in the Federal Government,” *International Journal of Electronic Government Research*, 2(1), 21-35 (Jan.-March 2006) as well as in a conference paper presentation at the Association for Public Policy Analysis and Management (APPAM) in November 2005.

7. REFERENCES

- [1] Fountain, J. E. *The Building of the Virtual State: Information Technology and Institutional Change*. Washington DC, Brookings Institution Press, 2001.
- [2] Fountain, J. E. Information, Institutions and Governance: Advancing a Basic Social Science Research Program for Digital Government.” Working Paper No. 03-004. Available at: http://www.umass.edu/digitalcenter/Research/working_papers/rwp03_004_fountain.pdf
- [3] Scott, J. Organizational knowledge and the intranet. *Decision Support Systems*, 23, 1 (1998), 3-17; Allcorn, S. Parallel virtual organizations: managing and working in the virtual workplace. *Administration and Society* 29 (1997), 412-39.
- [4] Scheepers, R. and Rose, J. Organizational intranets: cultivating information technology for the people by the people. In Subhasish Dasgupta (ed.) *Managing Internet and*

- Intranet Technologies in Organizations: Challenges and Opportunities*, Hersey, PA., Idea Group Publishing, 2002.
- [5] Peristeras, V., Tsekos, T. and Tarabanis, K. *Analyzing e-government as a paradigm shift*, (paper delivered at the Annual Conference of the International Association of Schools and Institutes of Administration, June 2002, The United Nations Thessaloniki Centre for Public Service Professionalism).
- [6] Martin, D., Rouncefield, M. and Sommerville, I. Applying patterns of cooperative interaction to work (re)design: e-government and planning. Proceedings of the SIGI conference on human factors. 2002. <http://www.dirc.org.uk/publications/inproceedings/papers/65.pdf> (10 Oct. 2005); Harvey, M., Palmer, J. and Speier, C. Implementing intra-organizational learning: a phased-model approach supported by intranet technology. *European Management Journal* 16, 3 (June, 1998), 341-354.
- [7] Fountain, Information, Institutions and Governance, p. 41.
- [8] Goldsmith, S. and Eggers, W. *Governing By Network*. Washington, D.C., Brooking Institution Press, 2004.
- [9] Brown, J. S. and Duguid, P. 1991. Organizational learning and communities-of-practice. *Organization Science*, 2,1 (February, 1991), 40-57.
- [10] Weick, K. *The Social Psychology of Organizing*. 2nd. ed. Reading, MA, Addison-Wesley, 1979.
- [11] Stacy, R.. 1995. The science of complexity: an alternative Perspective for strategic change processes. *Strategic Management Journal*, 16, 6 (September, 1995), 477-495; Drazen, R. and Sandelands, L. Autogenesis: a perspective on the process of organizing. *Organization Science*, 3,2 (May, 1992), 230-249.
- [12] Hayek, F. *Individualism and the Economic Order*. Chicago, University of Chicago Press, 1948.
- [13] Stacy, op.cit, p. 479.
- [14] Gil-Garcia, J. R. and Martinez-Moyano, I.J. *Exploring E-Government Evolution: The Influence of Systems of Rules on Organizational Action.*" NCDG Working Paper No. 05-001. Available at: http://www.ksg.harvard.edu/digitalcenter/Research/working_papers/gil-garcia_wp05-001.pdf (February 26, 2006).
- [15] Mahler, J. and Regan, P. M. The Evolution of web governance in the federal government. *International Journal of Electronic Government Research*, 2, 1 (January, 2006),
- [16] Mahler, J. and Regan, P.M. Crafting the message: controlling the content of agency websites. *Government Information Quarterly*, (forthcoming); Mahler, J. and Regan, P. M. Agency internets and changing dynamics of congressional oversight. *International Journal of Public Administration*, 28, 7&8 (2005), 553-565.
- [17] Lawrence, P. and Lorsch. J. *Organization and Environment*. Cambridge, MA, Harvard University Press, 1967.
- [18] Pfeffer, J. and Salancik, G. *The External Control of Organizations*. New York, Harper & Row, 1978.
- [19] Perrow, C. *Complex Organizations*. New York, Random House, 1986; Montjoy, R. and OToole, L. Toward a theory of policy implementation: an organizational perspective. *Public Administration Review*, (September/October, 1979): 465-476.
- [20] West, D. M. State and federal e-government in the United States, 2003. <http://www.insidepolitics.org/egovt03us.html> (25 Oct 2005).
- [21] West, D. M. State and federal e-government in the United States, 2005. <http://www.insidepolitics.org/egovt05us.pdf> (25 Oct., 2005)
- [22] Mahler, J. and Regan, P.M. Crafting the message.
- [23] Mahler, J. and Regan, P.M. Crafting the message.
- [24] Mahler, J. and Regan, P.M. Crafting the message.

Data Confidentiality, Data Quality and Data Integration for Federal Databases

Alan F. Karr

National Institute of Statistical Sciences
PO Box 14006
Research Triangle Park, NC, 27709-4006

karr@niss.org

1. OBJECTIVES AND IMPACT

The high-level goal of the research is to develop abstractions, theory and methodology and software tools that allow federal statistical agencies to disseminate useful information derived from confidential data but protect the privacy of data subjects—individuals and establishments.

Specific scientific objectives include problem formulations and scalable software tools that accommodate both *disclosure risk* and *data/information utility*, understanding *consequences of data integration* for data confidentiality, data quality, and creation of fundamental quantifications, usable models, scalable methods for *data quality*.

Impacts of the research include protecting government-collected data on individuals and establishments from increasingly severe threats to confidentiality, protecting privacy of individuals and establishments and helping agencies prepare for a possible “world without releasable microdata.”

2. SELECTED ACCOMPLISHMENTS

The project is leading to a paradigm shift in statistical disclosure limitation (SDL). New techniques are based on scalable risk-utility formulations that enable agencies to balance disclosure protection against the utility of released information.

Geographical Aggregation. Algorithms and software for achieving disclosability by aggregating adjacent geographical units such as counties were developed using data provided by the National Agricultural Statistics Service (NASS) [10, 11, 14]. These enable release of information below the state level, which was previously thought infeasible.

Tabular Data. NISS table servers [6] are the first implementation of query systems containing principled, scalable methods for dealing with query interaction. The project has also developed scalable methods and software to compute bounds on cell entries from released marginals as well as scalable risk-utility formulations [1, 2], and methods for releasing conditional distributions [3].

Data Swapping. Principal products of the research are complete risk-utility formulation for data swapping as a decision problem, with multiple measures utility/distortion and disclosure risk [5], the NISS Data Swapping Toolkit (DSTK) [15]—operational software for performing swapping large-scale studies to select the swap attributes and swap rate, as well as for visualization of the results, and a Web service implementation of data swapping [16].

Remote Access Analysis Servers. The project has created fundamental abstractions for systems that disseminate the results of statistical analyses of confidential data. *Regression servers* that optimally protect a sensitive variable have been developed [4], and software has been built.

Secure Analysis of Distributed Data. The project has developed methodology and implemented secure systems for data integration, construction of contingency tables whose cells contain either counts or sums, linear regression on horizontally and vertically partitioned databases and maximum likelihood estimation for exponential families [17, 18, 7, 12, 8, 13].

Framework for Comparing Statistical Disclosure Limitation Methods. A framework has been developed for comparison and combination of SDL methods for microdata. It incorporates multiple measures of data utility—ranging from very general to analysis-specific—as well as multiple measures of disclosure risk [9].

Research in Progress. Current research is addressing combining multiple SDL methods, which can yield improved disclosure risk and utility, as compared with either method alone, new measures of data utility based on distribution functions, clustering and propensity scores, all of which measure (in)ability to distinguish masked from original data, secure maximum likelihood estimation for general models, a new paradigm for data swapping, in which attributes to be swapped are also randomized, and secure regression for complex data partitions.

3. PROJECT STRUCTURE

The National Institute of Statistical Sciences (NISS) is lead institution for the project. University partners Carnegie Mellon University, Duke University, Iowa State University, Pennsylvania State University, Purdue University and Southern Methodist University. Federal statistical agency partners are the Bureau of Labor Statistics (BLS), Bureau of Transportation Statistics (BTS), Census Bureau (Census), NASS and National Center for Education Statistics (NCES). All have provided both data and support.

4. ACKNOWLEDGMENTS

This research was supported by NSF grant EIA-0131884 to NISS.

5. REFERENCES

- [1] A. Dobra, S. E. Fienberg, A. F. Karr, and A. P. Sanil. Software systems for tabular data releases. *Int. J. Uncertainty, Fuzziness and Knowledge Based Systems*, 10(5):529–544, 2002.
- [2] A. Dobra, A. F. Karr, and A. P. Sanil. Preserving confidentiality of high-dimensional tabular data: Statistical and computational issues. *Statist. and Computing*, 13(4):363–370, 2003.
- [3] S. E. Fienberg and A. B. Slavkovic. Bounds for cell entries in two-way tables given conditional frequencies. In J. Domingo-Ferrer and V. Torra, editors, *Privacy in Statistical Databases '2004*. Springer-Verlag, New York, 2004.
- [4] S. Gomatam, A. F. Karr, J. P. Reiter, and A. P. Sanil. Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access analysis servers. *Statist. Sci.*, 20(2):163–177, 2005.
- [5] S. Gomatam, A. F. Karr, and A. P. Sanil. Data swapping as a decision problem. *J. Official Statist.*, 21(4):1–21, 2005.
- [6] A. F. Karr, A. Dobra, and A. P. Sanil. Table servers protect confidentiality in tabular data releases. *Comm. ACM*, 46(1):57–58, 2003.
- [7] A. F. Karr, J. Feng, X. Lin, J. P. Reiter, A. P. Sanil, and S. S. Young. Secure analysis of distributed chemical databases without data integration. *J. Computer-Aided Molecular Design*, November, 2005:1–9, 2005.
- [8] A. F. Karr, W. J. Fulp, X. Lin, J. P. Reiter, F. Vera, and S. S. Young. Secure, privacy-preserving analysis of distributed databases. *Technometrics*, 2006. Invited paper; under review.
- [9] A. F. Karr, C. N. Kohnen, A. Oganian, J. P. Reiter, and A. P. Sanil. A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician*, 2005. Submitted for publication. Available on-line at www.niss.org/dgii/technicalreports.html.
- [10] A. F. Karr, J. Lee, A. P. Sanil, J. Hernandez, S. Karimi, and K. Litwin. Disseminating information but protecting confidentiality. *IEEE Computer*, 34(2):36–37, 2001.
- [11] A. F. Karr, J. Lee, A. P. Sanil, J. Hernandez, S. Karimi, and K. Litwin. Web-based systems that disseminate information but protect confidentiality. In W. M. McIver and A. K. Elmagarmid, editors, *Advances in Digital Government: Technology, Human Factors and Public Policy*, pages 181–196. Kluwer, Amsterdam, 2002.
- [12] A. F. Karr, X. Lin, J. P. Reiter, and A. P. Sanil. Secure regression on distributed databases. *J. Computational and Graphical Statist.*, 14(2):263–279, 2005.
- [13] A. F. Karr, X. Lin, J. P. Reiter, and A. P. Sanil. Secure analysis of distributed databases. In D. Olwell, A. G. Wilson, and G. Wilson, editors, *Statistical Methods in Counterterrorism*, Lecture Notes in Statistics. Springer-Verlag, New York, 2006.
- [14] J. Lee, C. Holloman, A. F. Karr, and A. P. Sanil. Analysis of aggregated data in survey sampling with application to fertilizer/pesticide usage surveys. *Res. Official Statist.*, 4:101–116, 2001.
- [15] National Institute of Statistical Sciences. Data Swapping Toolkit, 2003. Available on-line at www.niss.org/software/dstk.html.
- [16] A. P. Sanil, S. Gomatam, A. F. Karr, and C. Liu. *NISSWebSwap*: A Web Service for data swapping. *J. Statist. Software*, 8(7), 2003.
- [17] A. P. Sanil, A. F. Karr, X. Lin, and J. P. Reiter. Privacy preserving analysis of vertically partitioned data using secure matrix products. *J. Official Statist.*, 2004. Submitted for publication.
- [18] A. P. Sanil, A. F. Karr, X. Lin, and J. P. Reiter. Privacy preserving regression modelling via distributed computation. In *Proc. Tenth ACM SIGKDD Internat. Conf. on Knowledge Discovery and Data Mining*, pages 677–682, 2004.

Integrating Data and Interfaces to Enhance Understanding of Government Statistics: Toward the National Statistical Knowledge Network Project Briefing

Gary Marchionini

Stephanie Haas

University of North Carolina

100 Manning

Chapel Hill, NC 27516

011.919.966.3611

{march; haas}@ils.unc.edu

Catherine Plaisant

Ben Shneiderman

University of Maryland

Human-Computer Interaction Lab

College Park, MD

001.301.405.2680

{plaisant; ben}@cs.umd.edu

ABSTRACT

This paper reports the results of work in the final no-cost extension year of a digital government project that developed user interface models and prototypes to help people find and understand government statistics; proposed a Statistical Knowledge Network architecture that supports cross agency information access; and demonstrated models for government-academic collaboration.

Categories and Subject Descriptors

H.5 [[INFORMATION INTERFACES AND PRESENTATION](#)]

General Terms

Experimentation, Human Factors, Standardization

Keywords

User interfaces, metadata models, digital government, statistical information, information retrieval, online help

1. INTRODUCTION

This project aimed to help people find and understand government statistical information. We were especially interested in supporting non-specialist access to the substantial statistical data collected by various government agencies. Our research focal point was to create easy to use user interfaces that support both finding and making sense of pertinent statistical data that may exist in different government websites. This general focal point required that we develop user interfaces that citizens from all walks of life can use to achieve four goals: finding/exploring

statistical data; understanding the found statistical data and accompanying contextual information; helping information seekers use the systems and clarify the meanings of the data; and operate across government agencies.

In addition to the primary research and development challenges of building and evaluating novel user interface prototypes that achieved these goals, two distinct research challenges underpinned the interface work. These two foundational research threads are metadata and collaboration. In this briefing, we first outline our efforts to address these two foundational research threads, then summarize the interface work for each of the four goals, and conclude with reflections on the overall theme of integration. See the project website (<http://www.ils.unc.edu/govstat/>) for the more than 50 papers that have been published based on this work.

2. METADATA AND COLLABORATION

It is evident that high quality and expansive metadata is required to index corpora so that information can be found and contextualize search results so that found information can be understood. Additionally, we recognize that good metadata can be leveraged to provide online help services as people interact with the government websites. Likewise, standardized metadata is necessary for different computational systems to interoperate and there are a myriad of systems within a large government statistical agency as well as across agencies. In the early years of the project we worked to develop a statistical DTD for our agency partners' web-based statistical tables based upon DDI and NISO 11179 standards. This work led to the development of a layered model for metadata that agencies could use to add metadata to their sites. This work is ongoing in conjunction with Dan Gillman at the BLS.

As we worked to develop metadata models for statistical information, we were also developing user interfaces that required actual metadata for instantiation and testing. Our efforts to acquire large volumes of metadata from agencies

led to investigations into ways to automatically generate topical metadata. To this end, a research thread that used machine learning techniques to discover categories of webpages from crawled agency websites was undertaken. This work led to a text mining toolkit (available for download on the project website under demos and software) which was the basis for Elsas' Master's thesis. Examples of automatically generated topical categorizations for the entire website of each of our six government partner agencies have been produced.

In addition to the metadata work, we recognized that collaborative models were crucial to success. There were three kinds of institutional collaborations necessary: agency-agency, academic-academic, and academic-agency. We were extremely fortunate on all three types to have excellent bases upon which to build the trust necessary to move forward. The federal statistical community has long held cross-agency meetings and engaged in cross-agency projects, most notably for our project, the FedStats website that brought together data from all federal agencies that produce statistics. We continue to give talks and advice to cross agency groups (e.g., the National Infrastructure for Community Statistics). We were able to leverage this existing collaboration by conducting studies of the FedStats website and developing alternative interface options. Second, the collaboration between researchers at two different universities was fostered by long-standing and strong personal relationships grounded in other collaborative research. Third, we were able to quickly develop the academic-agency collaboration because there was also a history of collaboration among key agency personnel and the academic partners. Thus, we were able to develop a project team from the first months that engaged the attention of all participants and enabled us to have well-attended bi-annual all-project meetings, and active subteams that drew members from agencies and the universities.

These collaborations across institutions are of course rooted in personal relationships and this is both a great advantage and an Achilles heel for large-scale, long-term projects. In fact, due to a number of retirements in the agencies and graduations in the universities, much of the project momentum is lost. It seems that complex collaborative projects may have life-cycles that should be considered as projects mature.

3. USER INTERFACES

A variety of novel user interfaces have been developed and tested over the course of the past four years. One thread of work was devoted to developing and instantiating Relation Browser interfaces that support exploration of statistical websites across facets such as topic, time, geography, and data type. Over the course of the project, several user studies were conducted, the interface underwent a major

redesign, and was implemented for all of the agency partners' websites (see the project website for demos). Another set of user interfaces built upon treemap approaches that allow users to get an overview of large hierarchical data sets. Kules' dissertation demonstrated that meaningful and stable categories were effective organizing principles for search results. Another interface thread addressed the challenge of visually impaired users of maps. A sonification map browser was developed and evaluated in Zhao's dissertation. Each of these user interface projects aim to support exploratory search and to some extent, understanding results. The key remaining element for all of these interfaces is to link results to rich contextual metadata or help.

Another major effort for the project was to investigate and develop online help. An overarching multilayered framework for interface design that facilitates evolutionary learning and help was developed and illustrated with novel user interface prototypes. To directly address the issue of help for statistical concepts, an interactive statistical glossary was designed and tested and was the basis for Wilbur's Masters thesis. A project-led help symposium led to a general human information ecology model for how people find and understand information. Another thread of work was to develop narrated ShowMe! demonstrations for procedural operations with user interfaces.

4. CONCLUSION

Over the four years of this project, we have worked toward a vision of a generalized statistical knowledge network. We view user interfaces as the glue between the needs of people who want to use statistical information to make decisions and the statistical agencies that collect data on every aspect of human life. We have proposed a model based on tiers of data repositories available to online communities of citizens, organizations, and government agencies. We have developed and evaluated a set of novel user interfaces to support finding and understanding statistical information. We have also been among the leaders in developing new strategies for online help in the WWW environment. We have worked with agency partners to develop models for statistical metadata and processes for adding metadata to their collections, including the feasibility of leveraging automatic techniques. We have also demonstrated collaborative models for academic-government partnerships and note that to be pragmatic, such models should have a life cycle component.

5. ACKNOWLEDGMENTS

This work is supported by the National Science Foundation (NSF) under Grants EIA 0131824 and EIA 0129978 and from additional contracts from the Bureau of Labor Statistics, Census Bureau and National Center for Health Statistics.

SESSION 8C

INTERNATIONAL DIGITAL GOVERNMENT PROJECTS

Moderator

Jane Fountain, University of Massachusetts, Amherst, USA

Titles and Authors

Accelerated Indexing in a Domain-Specific Digital Library

Delcambre, Lois; Price, Susan; Nielsen, Marianne Lykke; Tolle, Timothy; Luk, Vibeke;
Weaver, Mathew

LOG-IN Africa: Local Governance and ICTs Research Network for Africa

Misuraca, Gianluca

Building efficiency through ICT utilization in the Government of Japan

Okumura, Hirokazu

Accelerated Indexing in a Domain-Specific Digital Library

¹Lois Delcambre, ¹Susan Price, ²Marianne Lykke Nielsen, ³Timothy Tolle,
⁴Vibeke Luk, ⁵Mathew Weaver

¹Computer Science Department
Portland State University
{prices, lmd}@cs.pdx.edu

⁴sundhed.dk
Copenhagen, Denmark
vlu@sundhed.dk

²Royal School of Library & Information
Sciences, Aalborg, Denmark
mln@db.dk

³Consultant
Vancouver, WA, USA
timtolle@aol.com

⁵Consultant
Paradise, Utah
mweaver@cs.pdx.edu

ABSTRACT

In this paper we summarize our progress and plans in two related, digital government research projects that focus on information retrieval in a domain-specific, digital library. We focus on supporting expert users in their various work tasks.

Categories and Subject Descriptors

H.3.1 [Information Storage and Retrieval]:Content Analysis and Indexing. H.3.3 - *Indexing methods* [Information Storage and Retrieval]:Information Search and Retrieval - *Search process*
H.3.7 [Information Storage and Retrieval]:Digital Libraries.

1. INTRODUCTION

The goal of this work is to improve the precision of information retrieval by exploiting domain-specific terminology, document types, and content. We adopt a hybrid approach to information retrieval where we combine high-quality, controlled vocabulary-based indexing with automatic text-based indexing because decades of research comparing the two techniques has not identified one technique as superior [1-3]. There is general recognition that intellectual, controlled indexing and automatic indexing should be used in combination.

Our research is motivated by two, distinct government domains: natural resource management and healthcare. We are exploring various mechanisms to improve searching by improving the quality and consistency of indexing, facilitating relevance judgments by the user, and improving ranking of search results.

2. FOREST PORTAL PROJECT

In our first project, we partnered with Region 6 of the USDA Forest Service in the US to provide access to documents produced as a part of the work processes of natural resource managers in a number of federal and other agencies. These

documents, including documents required by the National Environmental Protection Act (NEPA), represent the intellectual analysis, interpretation, and judgments of the interdisciplinary team involved in projects.

We focused on accommodating the many, many different controlled vocabularies, classification schemes, and terminologies used in the various scientific disciplines¹ (often with phrases that use ordinary English words with special meaning and where the same phrase can appear in multiple vocabularies). We developed a path-based thesaurus model and an associated digital library system, called Metadata++, with the following contributions: (1) representing multiple occurrences of terms with disambiguation based on the path, (2) supporting browsing documents associated with terms in the vocabularies, (3) calculating synonyms for geographic terms using GIS², and (4) supporting fast retrieval and vocabulary maintenance with very large hierarchies of terms.

Recent accomplishments include the completion of usability tests with eight Forest Service personnel from two National Forests. The users reported that they felt comfortable with the path-based representation of vocabularies, including multiple occurrences of specific terms [4-5]. Also, Weaver successfully defended his PhD dissertation [6].

3. DANISH HEALTH PORTAL PROJECT

In our current project, we are working with documents from natural resource management and also with documents in the operational, national Health Portal in Denmark (sundhed.dk). The Health Portal is intended to provide information on health and the healthcare system to citizens and clinicians. Of the nearly 22,000 documents in the Health Portal, where documents are manually indexed using terms from one or more of the three (Danish) vocabularies in the system. **Almen** is a home-grown thesaurus with 1,477 terms that is intended for use by non-professionals. **ICPC** is an international classification scheme intended for use by clinicians with 712 terms and **ICD-10** is a detailed, international vocabulary with 18,441 terms. Of the 22,503 terms, 317 terms appear twice, 28 terms appear three

¹ NSF Grant Number 9983518, Digital Government Program, August 2000 – March 2005.

² GIS is a geographic information system, capable of spatial and geographic reasoning.

times, and 5 terms appear more than three times. The term “Funktionsindskrænk/handicap INA³” appears the most often (16 times). These vocabularies mainly have terms for “disease or syndrome,” which is important, but few terms that describe other important concepts in healthcare, such as terms that describe tests or procedures. We are currently considering whether we might incorporate these additional vocabularies into the portal to test our path-based thesaurus system.

We have also considered the work tasks and the associated information needs of family physicians⁴ that use the portal. This has led us to define the notion of “semantic components” for particular, domain-specific document classes [7]. For example, a physician seeing a patient with puzzling symptoms might be interested only in documents with information about *diagnosis* of migraine headaches, not about *treatment* or *prevention*. The italicized words are semantic components. Work to date has confirmed that documents in the Health Portal as well as forest documents can be classified into distinct classes, where each class has associated semantic components. More than that, the semantic component instances are not necessarily demarcated by document structure. We are investigating the following research questions. (1) Can information needs be usefully expressed as full-text search within specifically-named semantic components? (2) (Given that information needs can be usefully expressed,) will retrieval based on such searches improve precision? (3) (Given that precision is improved,) can we automate or semi-automate the identification of semantic components? (4) Is it easy to index documents by (only) indicating the semantic component instances? (5) Is indexing documents by (only) indicating the semantic component instances more consistent than typical keyword indexing? And, finally, (6) how can we quantify the performance of information retrieval using semantic components?

4. DISCUSSION

Our two government settings differ in that the Danish Health Portal is an operational, web-based portal used across Denmark whereas the USDA Forest Service does not have an implemented portal. This may make it easier (or harder!) to transfer results from our research to the operational setting in Denmark.

One challenge of working with the Danish Health portal is that the vocabularies and the documents are in Danish. However, our Danish colleagues as well as our collaborators at sundhed.dk are fluent in English. Price (PhD student working on this project) decided to learn a bit of Danish. Given that she is a domain specialist (an MD), it was somewhat easy for her to understand medical terminology in Danish.

We are currently planning usability tests in Denmark in the Summer of 2006, with financial support from the Danish Region

³ The English version of this ICPC term is “General and unspecified limited function/disability NOS” where NOS means “not otherwise specified.”

⁴ This work is being conducted in collaboration with Peter Vedsted, a physician in Denmark who has been involved in the development and evaluation of the Health Portal.

of Århus (~\$9,000) to compensate family physicians who participate in our tests.

We are working on automating the recognition of semantic components in forest documents, in part, because they are in English and because their content is prescribed by NEPA.

Government-based digital libraries offer a rich environment for our research; the documents and the work tasks are inherently domain-specific. Access to expert users to learn about their work tasks and information needs is invaluable. The presence of multiple jurisdictions, with some autonomy at each level, is typical in government settings. The settings for our research are not different in this regard; both environments are governed by competing national, regional, and local issues and concerns. This can make it harder to achieve consistency in procedures which can in turn make it harder to improve information retrieval.

5. ACKNOWLEDGMENTS

This work is supported in part by the National Science Foundation, grant number 0514238. Any opinions, findings, conclusions, or recommendations expressed here are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

6. REFERENCES

- [1] Anderson, J. and Pérez-Carballo, J. The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part I: Research, and the nature of human indexing and Part II: Machine indexing, and the allocation of human versus machine effort. *Information Processing & Management*, 37, pp. 231-277, 2001.
- [2] Rowley, J. The controlled versus natural indexing languages debate revisited: a perspective on information retrieval practice and research. *Journal of Information Science*, 20(2), pp. 108-119, 1994.
- [3] Savoy, J. Bibliographic database access using free-text and controlled vocabulary: an evaluation. *Information Processing & Management*, 41, pp. 873-890, 2004.
- [4] Weaver, M., Delcambre, L. M. L., Tolle, T. and Nielsen, M. L. Using a Path-Based Thesaurus Model to Enhance a Domain-Specific Digital Library. In *Proceedings of the European Conference on Digital Libraries (ECDL) 2005*, pp. 511-512, Vienna, Austria, 2005.
- [5] Nielsen, M., Delcambre, L., Tolle, T. and Weaver, M. Indexing and retrieval challenges in digital government systems - summary of an empirical research project. In *2nd Scandinavian Workshop on eGov*, Copenhagen, Denmark, 2005.
- [6] Weaver, M., Enhancing a domain-specific digital library with metadata based on hierarchical controlled vocabularies. PhD Dissertation, Oregon Health & Science University, 2005. L. Delcambre, Advisor.
- [7] Price, S., Delcambre, L., Nielsen, M. L., Tolle, T. Luk, V. Weaver, M., “Using Semantic Components to Facilitate Access to Domain-Specific Documents in Government Settings”, Proc. of the 2006 National Digital Government Conference (dg.o 2006), 2 pages.

LOG-IN Africa

Local Governance and ICTs Research Network for Africa

Gianluca Misuraca
Project Leader / Coordinator
CAFRAD, +212 61 307269
misuraca@caftrad.org

LOG-IN Africa is an emergent pan-African network of researchers and research institutions from nine countries (Egypt, Ethiopia, Kenya, Mauritius, Morocco, Mozambique, Senegal, South Africa, and Uganda). It will assess the current state and outcomes of e-local governance initiatives in Africa, and in particular how ICTs are being used to realise good local governance at four levels: a) the internal organizational processes of local governments; b) the provision of information and service delivery; c) the promotion of the principles of good governance; and d) public participation and consultation. As a network of research peers, it will be coordinated through the African Training and Research Centre in Administration for Development (CAFRAD), ensuring effective implementation; a pan-African outlook; and high-level policy dissemination of research results.

Through collaboration among the networked research partners, a modular “outcome assessment framework”, with specific indicators linked to a conceptual framing of the characteristics of “good governance”, will be developed. This will be adapted to local contexts in conjunction with training of local researchers. Specific data collection and analysis will be conducted at the local level in large part based on locally determined research priorities, strategies and methods broadly coordinated within the LOG-IN Africa research framework and research objectives. An integrative process will allow for comparative cross-national and regional assessment of the outcomes of current e-local governance activities.

The Network will generate research findings contributing to more effective policy making and implementation in e-local governance in Africa. National and regional guidelines and an implementation “Roadmap” for how to proceed in this area will also be developed. The Network research partners will be drawn from the following institutions:

1. Al Akhawayn University, Morocco (<http://www.aui.ma>)
2. Cairo University, Egypt (<http://www.cu.edu.eg>);
3. LINK-Centre/Witswatersrand University and the Centre for Public Service Innovation (CPSI), South Africa (<http://link.wits.ac.za> and www.cpsi.co.za);
4. Makerere University, Uganda (<http://www.cit.ac.ug>);
5. Société Africaine d'Education et de Formation pour le Développement / African Society of Education and Training for Development (SAFEFOD), Senegal (www.safefod.org);
6. University of Addis Ababa, Ethiopia (<http://www.aau.edu.et>);
7. Universidade Eduardo Mondlane, Mozambique (<http://dmi.uem.mz>);
8. University of Nairobi, Kenya (<http://www.uon.ac.ke>);
9. University of Technology of Mauritius, Mauritius (<http://ncb.intnet.mu/utm>).

The research Project LOG-IN Africa, has the general objective of informing, supporting and orienting African countries and other stakeholders in their policies and practices concerning the application of ICTs to local governance.

The specific objectives of the Project are as follows:

- 1) To establish an “open” pan-African research network, based on key partner institutions, to support research, training and implementation of activities in the area of ICTs for local governance in Africa
- 2) To define a common framework and an accepted methodology of measurement; and to assess the multi-dimensional effects of ICTs on local governance in Africa.
- 3) To develop guidelines for “e-Local Governance” projects in Africa in support of the successful local implementation of ICTs for local governance.
- 4) To increase awareness and reinforce institutional capacity in ICTs for local governance in Africa; and,
- 5) To widely disseminate its research results in order to influence policies, inform practice and add to the knowledge base in the area of e-Local Governance.

During the two-year of its implementation (January 2006-January 2008), the Research Network is intended to contribute to improved project management and evaluation capabilities of the participating institutions.

The direct outputs of the research project, the outcome assessment framework and the Road Map for e-Local Governance, will shape the landscape in the theory and practice of ICTs for local governance. The establishment of the Network will also facilitate the gaining of support for e-local governance from the international community as for example, through the framework of the WSIS and the recently inaugurated Digital Solidarity Fund (DSF). The Network will serve as a possible channel for resources from the North to the South, in this area as well as supporting South-South partnerships, and decentralized cooperation. A particular emphasis will be put on the leveraging role of ICTs in supporting local economic development through e-Local Governance activities as for example, promoted by the African Diaspora and through local-to-local connections (e. g. cooperation among Regions and Cities).

In the longer term, in addition to reinforcing the research capacities of partner institutions in the area of ICTs and local governance and facilitating the sharing of knowledge and experience, the Research Network will reinforce the capacities of local governance institutions and grass-roots communities. In this way the project will potentially have an additional impact through building the skills of young leaders and ICTs professionals, thus supporting the development of human resource capacity at the local level for the implementation of ICTs projects and in turn creating job opportunities and stimulating economic growth.

Building efficiency through ICT utilization in the Government of Japan

Hirokazu Okumura

Visiting Professor, University of Tokyo

Visiting Research Fellow,

National Center for Digital
Government, Spring 2005

(in association with J. E. Fountain)

[jokumura@mail.ecc.u-
tokyo.ac.jp](mailto:jokumura@mail.ecc.u-tokyo.ac.jp)

ABSTRACT

In this paper, we describe the legislative, institutional and operational developments regarding electronic government initiative in Japan as a first step of U.S. Japan Comparative Digital Government Research.

General Terms

Management of electronic government initiatives

Keywords

Governance, management in the government, technology enactment theory, bureaucratic actors' behavioral analysis

1. INTRODUCTION

The central government of Japan has worked on the development of electronic government since the 1990s. After 2001, the effort has been accelerated. This paper follows recent movement of the Japanese Government towards the most advanced Information and Communication Technology (ICT) nation.

Provisional hypotheses are presented from the perspective that understands the reciprocal influences between the bureaucracy and ICT based on "technology enactment theory" with actors involved in the development process of the policy.

2. ELECTRIC GOVERNMENT POLICY OF THE JAPANESE GOVERNMENT AFTER 2001

1.1 Enactment of "Basic Law on the Formation of an Advanced Information and Telecommunications Network Society"

The law was enforced in January, 2001, and it stipulates (a) the basic principles and ideas for developing policy measures, (b) clarifications of the responsibilities of the central and local governments, (c) establishment of the Strategic Headquarters for the Promotion of Advanced Information and Telecommunications Network Society in the Cabinet, and (d) the obligation for the central government to decide (annual) priority policy programs for the development of Advanced Information and Telecommunications Network Society".

1.2 Institutional development of the promotion of electronic government

The Agency Chief Information Officer (CIO) in each agency was created by the decision of "Government Administration Informatization Promotion Council" on 30 July, 2002. In addition, a CIO council was established by the order issued by the Prime Minister, Chief of the Strategic Headquarters for the Promotion of an Advanced Information and Telecommunications Network Society.

Agency CIOs in the Japanese Government are not experts on information technology unlike those in private firms and in the United States federal government. Rather they are top level officials in charge of business execution of the agencies. Agency CIOs have no strong incentives to rethink and manage IT as a tool for performance enhancement of administration and usually they do not have enough training or business experience to absorb advice from Technical Advisors to Agency CIOs.

Moreover, the CIO council decided to hire "experts who have special knowledge and experience concerning business process analysis, information systems technology, and information security" as Technical Advisor(s) to the

Agency CIO in every agency to cover both their lack and lagging of ICT expertise.

Technical Advisors to Agency CIO remain as advisors and do not have executive powers.

1.3 Electronic Government Development Plan

The plan to develop Electronic Government comprehensively was decided at the CIO Council in July 2003 after several months of negotiations among all agencies. It has two targets: (a) improvement of convenience and services to citizens, and (b) governmental business process reform corresponding to IT. The planning period is 3 years from fiscal year 2003 to the end of fiscal year 2005. A new method adopted by the Japanese Government is the "Business and System Optimization Program", which is a Japanese version of Enterprise Architecture in the Government.

All agencies in the government are now deliberating to complete their "Business Process and System Optimization Program" including 14 across the agency programs.

2. VISUALIZATION OF BUSINESS PROCESS IN THE GOVERNMENT

By analyzing the business process, we have visualized the "AsIs" process that 35 staff, managers and employees are involved in during the purchase decision-making process of commodities in an agency. It takes about one month before goods are obtained. Then, what would be a desired future state, called "ToBe," of the acquisition process? The ToBe process has not yet been well developed. But, an indispensable business process would have only four routines: (a) judgment of purchase of goods, (b) check on limitation of budget, (c) selection and contract of cheap supplier, and (d) actual expense from the budget. It seems out of date and redundant that as many as 35 staff are engaged in the buying process of goods.

3. PROVISIONAL HYPOTHESES FOR FUTURE RESEARCH

Seven hypotheses below are advanced for future examination. They originate from exploratory interviews and my thirty years experience as a government official.

1) Information system experts in agencies who carry out the business architecture analysis of AsIs and ToBe do not and cannot keep in mind so much simplification and streamlining duplications of the business flows. Although a Technical Advisor to the Agency CIO tends to keep such business reform in mind, their power over implementation of the reform is very limited because of their position as an advisor.

2) The rationale for resistance to change towards the ToBe state might be that the present business process is best developed and has no serious problems. In other cases, business specificity (customization) is used as the reasoning against change.

3) Resistance could also be a rationale for protection of the present organization.

4) The cause of resistance might be the fear of the loss of employees' positions and reassignment to new and ill-experienced positions. The information system experts in an agency develop the information systems by supporting employees' standpoint.

5) The CEO also usually defends regular employees' point of view. The government doesn't have an incentive to create a desired future state, ToBe, leading to decreasing the number of employee and a drastic reshuffle as long as the effect of the IT investment cannot be foreseen clearly. (And, furthermore, the future effect on the government can not be measured easily).

6) If the business is more critical for an agency, the resistance of the agency to the utilization of the use of the integrated system between agencies or other agency systems/rules might be stronger.

7) An agency CIO and CEO might not understand and execute their powers to improve organizational performance by exploiting IT. And they also might not understand the importance of an AsIs visualization analysis and the business process reform that should be the first step before the development of an information system.

4. CONCLUSION

The author suggests that technology enactment theory clearly works also in Japan. Regarding the relations between IT and institutions, I would analyze various aspects of the relations between IT and institutions such as the general relations between IT and institutions in organizations, unique points in the government bureaucracy, and questions indigenous to Japan.

5. REFERENCES

- [1] A U.S.-Japanese comparative study was launched, a collaborative effort between the University of Tokyo, COE on Comparative Policy Analysis in Advanced Countries, and the University of Massachusetts Amherst, National Center for Digital Government under the supervision of the PI, Jane Fountain, and the author. The project highlighted here presents the most recent phase of this comparative project
- [2] H. Okumura, "Can Technology Promote Innovation in Japanese Government?" NCDG seminar paper, June 2005
- [3] J. E. Fountain, "Prospects for the Virtual State," Center of Excellence Program on Invention of Policy Systems in Advanced Countries, Graduate School of Law and Politics, University of Tokyo, working paper, Sept 2005
- [4] J. E. Fountain, "Building the Virtual State" Japanese translation by H. Okumura (Tokyo: Ichigeysha, 2005)
- [5] http://www.kantei.go.jp/foreign/it/it_basiclaw/it_basiclaw.htm
- [6] http://www.kantei.go.jp/foreign/policy/it/index_e.html

SYSTEM DEMONSTRATIONS

DURIAN: A Demo for Near-Duplicate Detection

Hui Yang

Language Technology Institute
School of Computer Science
Carnegie Mellon University
huiyang@cs.cmu.edu

Jamie Callan

Language Technology Institute
School of Computer Science
Carnegie Mellon University
callan@cs.cmu.edu

Stuart Shulman

Library and Information Science
School of Information Sciences
University of Pittsburgh
shulman@pitt.edu

ABSTRACT

Recently, the move from paper to electronic public comments makes it much easier for individuals to customize form letters while harder for agencies to identify substantive information since there are many near-duplicate comments that express the same viewpoint in slightly different language. The identification of exact- and near-duplicate texts, and recognition of unique text within near-duplicate documents, is an important component of data cleaning and integration processes for eRulemaking.

This brief paper describes a demonstration of a near-duplicate detection system, DURIAN (DUllicate Removal In lARe collectionN), that identifies and organizes the near-duplicates for eRulemaking applications.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval] Clustering.

General Terms

Algorithms, Performance, Experimentation.

Keywords

Duplicate detection, clustering, eRulemaking, text analysis.

1. SYSTEM OVERVIEW

DURIAN (DUllicate Removal In lARe collectionN) [1] is a system to detect and organize exact- and near-duplicates in notice and comment rulemaking. Near duplicates are generated in large volume when an advocacy organization distributes a form letter by email or posts the form letter on a web page that allows or encourages customization before submission. A regulatory agency would likely want these comments grouped together even though the amount of modified text varies greatly. Our goal in this work is to greatly improve near-duplicate detection accuracy for notice and comment rulemaking as well as to maintain efficiency.

Our research was conducted with two public comment datasets. One dataset, for the EPA's Proposed National Emission Standards for Hazardous Air Pollutants for Utility Air Toxics rule (USEPA-OAR-2002-0056, "Mercury rule"), contains 536,975 email messages. The second dataset, for the DOT's Proposed Average Fuel Economy Standards for Light Trucks rule (USDOT-2005-22223), contains 45,979 comments, including a mixed format of email messages, pdf files and scanned image files. Experimental evaluation of DURIAN's accuracy requires human assessment, which is impractical for the full datasets. Manual evaluation of near-duplicate detection accuracy on samples of each dataset

shows that system-human intercoder agreement is comparable to human-human intercoder agreement [Yang, et al, 2006].

The figure shows a screenshot of the DURIAN software interface. On the left, a list of 'Message Groups' is shown with their counts: 92-16-2004-009534 (110), 06-22-2004-520077 (86), 02-20-2004-134023 (46), 05-08-2004-108712 (42), 06-12-2004-513426 (31), 04-10-2004-371752 (23), 05-14-2004-501007 (21), 03-20-2004-175227 (19), 02-29-2004-153358 (11), 04-23-2004-400538 (9), 06-17-2004-511742 (9), 05-25-2004-505382 (7), 06-26-2004-529546 (7), 04-26-2004-413950 (5), 03-13-2004-163779 (5), 03-29-2004-354292 (5), 03-04-2004-156742 (4), 03-31-2004-360383 (3), 04-23-2004-405656 (2), and 07-11-2004-536083 (2). The middle pane displays message metadata for selected groups, and the right pane shows the raw message content with highlighted modified parts.

Message Groups				
group	size	subject	sender	delivered
92-16-2004-009534	110		-	getactive
06-22-2004-520077	86		-	getactive
02-20-2004-134023	46		-	getactive
05-08-2004-108712	42		-	getactive
06-12-2004-513426	31		-	getactive
04-10-2004-371752	23		-	getactive
05-14-2004-501007	21		-	getactive
03-20-2004-175227	19		-	getactive
02-29-2004-153358	11		-	getactive
04-23-2004-400538	9		-	getactive
06-17-2004-511742	9		-	getactive
05-25-2004-505382	7		-	getactive
06-26-2004-529546	7		-	getactive
04-26-2004-413950	5		-	getactive
03-13-2004-163779	5		-	getactive
03-29-2004-354292	5		-	getactive
03-04-2004-156742	4		-	getactive
03-31-2004-360383	3		-	getactive
04-23-2004-405656	2		-	getactive
07-11-2004-536083	2		-	getactive

MESSAGE 04-03-2004-363533 - [SHOW DIFF WITH SEED DOCUMENT IN NEW WINDOW]

We would not think of changing LEAD from a toxic substance to a non-toxic substance, why therefore, would we wish to change mercury, a similarity and just as lethal toxic? IT DOES NOT MAKE SENSE!

Mercury is a potent neurotoxin that can cause brain damage, especially to fetuses and young children. Studies have also shown a correlation between mercury levels and Alzheimer's, Parkinson's and other neurological diseases.

I am concerned that 1 out of 12 women in the United States have mercury in their blood at unsafe levels which means that each year hundreds of thousands fetuses are at risk of exposure to this toxic pollutant. We need strong rules regulating mercury emissions from the power sector now, to

Figure 1: The DURIAN software.

Figure 1 shows a DURIAN window. DURIAN is web-based and publicly accessible with password protection. The system identifies reference copies of form letters, modified copies of form letters (near-duplicates) and how they were modified, and unique comments. The left pane lists the document IDs of reference copies. When a reference document ID is clicked, the upper right pane shows message metadata, such as, subject, author, submitted date, and relayer information. The lower right pane shows the message contents and highlights the modified part of a form letter.

ACKNOWLEDGMENTS

We thank the DOT and EPA for providing the public comments that made this research possible. This research was supported by NSF grant IIS-0429102. Any opinions, findings, conclusions, or recommendations expressed in this paper are the authors', and do not necessarily reflect those of the sponsor.

REFERENCES

- [1] Hui Yang , Jamie Callan, Stuart Shulman,"Next Steps in Near-Duplicate Detection for eRulemaking", In *Proceedings of the 6th National Conference on Digital Government Research (DG.O2006)*, San Diego, California , May 21-24 2006.

webADIS: A Flexible web-based Environment for the Automated Dental Identification System

Satyasrinivas Chekuri, Diaa Eldin Nassar, Ayman Abaza, Eyad Haj Said, Ali Bahu,
Uthman Qurashi, Gamal Fahmy and Hany Ammar
Lane Dept. of Computer Science and Electrical Engineering
West Virginia University, P.O. Box (6109)
Morgantown, WV 26506-6109 – USA
{satyac, dmnnassar, ayabaza, hajsaid, alib, uqureshi, fahmy, ammar}@csee.wvu.edu

ABSTRACT

Automating the process of postmortem (PM) identification of individuals using dental records is receiving increased attention. In developing a research prototype of an Automated Dental Identification System (ADIS), research teams from multiple institutions collaborated with forensic experts from the US Federal Bureau of Investigation Criminal Justice Information Services division to identify the functional requirements of ADIS. A multitude of digital image processing and pattern recognition techniques were developed to meet the requirements of the constituent components of ADIS. In this demo, we present a web-based environment called webADIS that integrates ADIS components and provides a unified web-based interface. webADIS provides support for configuring the system into multiple possible realizations for some components as well as alternative identification strategies.

INTRODUCTION

Architecturally, ADIS consists of the following components: (i) the Record Preprocessing component, (ii) the Potential Matches Search component, and (iii) the Image Comparison component[1].

The record preprocessing component handles the following tasks: (a) cropping of records into dental films [2], (b) enhancement of films to compensate for possible low contrast, (c) classification of films into bitewing, periapical, or panoramic views [3], (d) segmentation of teeth from films [3]-[7], and (e) annotating teeth with labels corresponding to their locations[8].

The potential matches search component manages archiving and retrieval of dental records based on high-level dental features (e.g. number of teeth and their shape properties) and produces a candidate list. Several possible approaches to realize potential matches search are presented in [9].

The image comparison component is based on low-level tooth-to-tooth comparison between the subject record's teeth, after alignment [10] with the corresponding teeth of each candidate, thus producing a short match list.

Each task defines a subcomponent that has specific interfaces, which may be implemented using different

realizations. As an example, the teeth segmentation subcomponent, which achieves the preprocessing step of identifying the extent of teeth comprised in a digitized dental radiographic film, has been implemented using five different algorithms as described in [3]-[7].

webADIS: The Flexible Integration web Environment

In developing an environment that integrates the various ADIS components, we adopted a 3-tier architecture. In this architecture, volumes of data are maintained at a database layer, business rules (and logic) are processed (and validated) at the server or business tier, which also handles transactions with the client or presentation layer (the front end) [11]. There are several benefits for using the 3-tier architecture in developing an environment like webADIS:

- First, the middle (or server) tier provides a layer of abstraction and hence client applications do not depend on the type of database server that actually stores the data. So any changes made to the database-tier would only require changes in middle tier.
- Second, a component in the business layer can be accessed by any number of components in presentation layer (clients), therefore any component implementation changes made at the business-tier will not mandate changes in components of other two tiers.
- Third, this decoupling of tiers according to roles facilitates dynamic load balancing of servers, thus a server shares its load with other servers at times of bottlenecks.
- Finally, 3-tier implementations can be designed to make business objects and data storage so close (in a network sense) that the network load is eliminated, thus reducing inefficient utilization of network bandwidth.

Figure 1 illustrates the 3-tier architecture of webADIS, where a web browser component at the client tier communicates with a web server component at the middle tier, a central component (ADIS Server) handles transaction dispatching to the respective server components as well as to the database tier as needed.

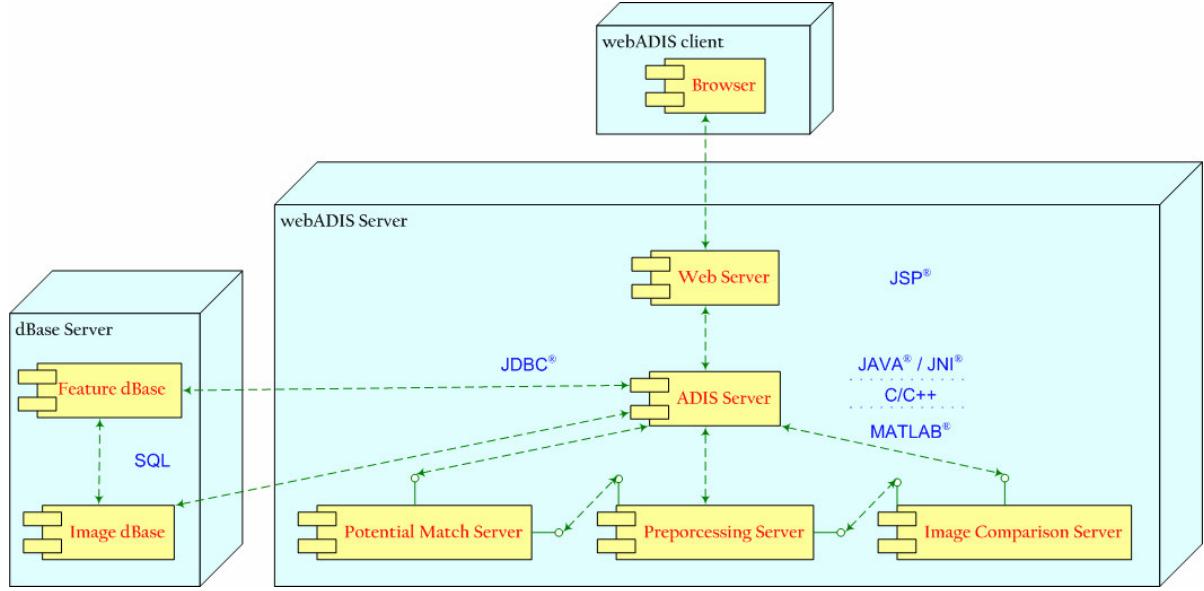


Figure 1: Architecture of webADIS; the web accessible ADIS environment.

Conclusion and Future Work

We presented a web-based environment (webADIS) that integrates the components of a research prototype of an Automated Dental Identification System (ADIS). This environment provides a flexible remote access to ADIS from virtually any computing platform connected to the Internet. Flexible remote access to forensic systems is important to law enforcement and digital government officers managing mass disasters (e.g. Tsunami). The webADIS environment facilitates seamless configuration of ADIS by so user can choose among multiple possible realizations of some components as well as alternative identification strategies.

In our future work we will focus on improving the timeliness of the application by considering different high performance computing paradigms, we will also continue to improve on the accuracy of the underlying image processing and pattern recognition algorithms.

Acknowledgement

This work is supported in part by the U.S. National Science Foundation under Award number EIA-0131079 to West Virginia University, the research is also supported under Award number 2001-RC-CX-K013 from the Office of Justice Programs, National Institute of Justice, U.S. Department of Justice. Points of view in this document are those of the authors and do not necessarily represent position of the U.S. Department of Justice.

References

- [1] Fahmy, G., Nassar, D., Haj Said, E., Ammar, H., Abdel-Mottaleb, M., Chen, H., Jain, A., "Toward an Automated Dental Identification System", *Journal of Electronic Imaging*, Vol. 14, No. 4, 2005.
- [2] Li, X., Abaza, A., Nassar, D., and Ammar, H. "Fast and Accurate Segmentation of Dental X-ray Records", *Proc. of 2006 International Conference on Biometric Authentication (ICBA)*, Hong Kong, Jan 2006.
- [3] Zhou, J., and Abdel-Mottaleb, M. "A Content-based System for Human Identification based on Bitewing Dental X-Ray Images", *Pattern Recognition*, Vol. 38, No. 11, pp. 2132-2142, 2005.
- [4] Anil K Jain and Hong Chen, "Matching of Dental X-ray images for Human identification", *Pattern Recognition*, vol. 37, pp. 1519-1532, 2004.
- [5] O. Nomir and M. Abdel-Mottaleb, "A system for Human Identification from X-Ray Dental Radiographs", *Pattern Recognition*, vol. 38, pp. 1295-1305, 2005.
- [6] Eyad Haj Said, Diaa Nassar, Gamal Fahmy, and Hany Ammar , "Teeth Segmentation and Enhancement in Digitized Dental X-Rays Films", to appear in *IEEE Transactions on Information Forensics and Security*.
- [7] Eyad Haj Said, Diaa Nassar, and Hany Ammar, "Image segmentation for automated dental identification", *Proc. of SPIE Electronic Imaging*, Jan 2006.
- [8] Mahoor, M., and Abdel-Mottaleb, M. "Automatic Classification of Teeth in Bitewing Images", *Proc. of ICIP 2004*, Singapore, August 2004.
- [9] H. Chen and A. Jain, "Dental Biometrics: Alignment and Matching of Dental Radiographs", *IEEE Transactions on PAMI*, Vol. 27, No. 8, pp. 1319-1326, August 2005.
- [10] Nassar, D., Ogirala, M., Adjeroh, D., and Ammar, H., "An Efficient Multi-Resolution GA approach to Dental Image Alignment", *Proc. of SPIE Electronic Imaging*, Jan 2006.
- [11] SatyaSrinivas Chekuri "A Web-Based Environment for Automated Dental Identification Research" Masters Thesis. Department of Electrical and Computer Engineering – WVU, 2005.

Living on the Edge with the Oregon Coastal Atlas

Paul Klarin, Tanya Haddad

DLCD/Ocean-Coastal Mgmt. Program

635 Capitol St. NE, Suite 150

Salem, OR 97301-2540

1-503-373-0050, ext. 249

paul.klarin@state.or.us,
tanya.haddad@state.or.us

Joseph Cone

Oregon Sea Grant

Oregon State University

Corvallis, OR 97331-2131

1-541-737-0756

joe.cone@oregonstate.edu

Dawn J. Wright

Oregon State University

Dept. of Geosciences

Corvallis, OR 97331-5506

1-541-737-1229

dawn@dusk.geo.orst.edu

ABSTRACT

In this paper, we describe an educational DVD entitled *Living on the Edge: Building and Buying Property on the Oregon Coast*, intended to alert homeowners, buyers, developers, realtors to the hazards associated with storms and other natural processes on the Oregon Coast. Understanding of these hazards and their impacts often relies upon geospatial data clearly communicated as a map. The DVD is a public outreach extension of an original Digital Government project, the Oregon Coastal Atlas, an interactive map, data, and metadata portal for coastal resources managers and scientists that was introduced at dg.o 2004

Categories and Subject Descriptors

D.4.0 [Operating Systems]: Microsoft Windows, Linux.

General Terms

Management, Documentation, Economics, Human Factors.

Keywords

Natural hazards, public education, geospatial data, Internet map servers, state government educational DVD, coastal resource management, coastal GIS, web GIS, atlas.

1. INTRODUCTION

The Oregon Coastal Atlas (OCA) is an interactive map, data, and metadata portal for coastal resources managers and scientists that was introduced at dg.o 2004 [5]. The State of Oregon's Ocean-Coastal Management Program (OCMP) in Portland and Salem, Oregon State University in Corvallis, and Ecotrust, a non-profit organization in Portland, developed it. The OCA has addressed information technology research issues and potential solutions for the benefit of *decision-makers*, at the level of state and local government and in various non-governmental organizations [1]. Indeed, there are still many challenges faced by these practitioners, including gaps in data, effective data integration, data presentation, how to turn existing *data* products and information management tools into useful *information* products, and how to use or create appropriate indicators of varying types (e.g., hazard, health, suitability, etc.). In Oregon, effective coastal management relies largely on the outcome of resource decisions made at the local level, by local officials and ordinary citizens [4]. The prior NSF Digital Government (DG) award that established the OCA, has shown that collaboration between a state agency and a team of academics and private sector scientists has been crucial in developing the *initial* computational infrastructure need for providing this

information at the state and local level. The OCA has been unique in that it couples up-to-date, interdisciplinary resource data along with accompanying spatial analysis tools that can be used either online within the OCA or downloaded to the desktop.

We now seek to extend this effort to the realm of informal science education. Toward this end, a DVD has been developed (*Living on the Edge: Building and Buying Property on the Oregon Coast*), with excerpts to be prepared for inclusion in the "Learn" section of the OCA, which provide background information on "Coastal Systems" such as estuaries, sandy shores, rocky shores, and ocean areas, "Coastal Topics" such as hazards, public access, fisheries, processes, etc., and cross-links to relevant datasets available in the OCA's GIS archive. The primary audience now shifts from coastal resource managers to new or prospective residents, property owners, developers, lenders, insurers and real estate professionals interested in residing, purchasing and developing property along the Oregon coast. The objective of the DVD is to educate about the dynamic nature of the geologic and climatic settings prevalent along the coast, and to encourage more informed decision-making. The DVD includes comparative summer/winter footage of beaches experiencing erosion and storm events, interviews with scientists, experts, planners, property owners, advice on where to get information and what kinds of questions to ask regarding a site. Scientific topics that are addressed include bluff failures (for which interactive tools are already available within the OCA), coastal erosion/deposition, tsunamis, landslides and flooding. The DVD is now available from Oregon Sea Grant Communications, with target distribution to local cable channels, libraries, community colleges, realtor groups, rotary and other clubs, and neighborhood associations.

2. IMPACT

Oregon's coastline is a beautiful, dynamic and exciting environment, offering dramatic landscapes and bountiful resources. But the same elements that make it so attractive are also responsible for a variety of natural climatologic and geologic hazards, both chronic and catastrophic that can affect both people and property. These hazards are active over a wide temporal and spatial range. Some, like seasonal storms and erosion are continual and predictable. Others like earthquakes and tsunamis, are rare but potentially devastating. All of them pose a risk that can be mitigated to some degree. History has shown that those who were ignorant of the dangers or ignored them, have paid a steep price, at sites such as Jump Off Joe and Bayocean spit, Nesika [3]. The more one knows about the natural processes affecting a particular parcel of land, the better off one is in making decisions about buying, building

on, or using it. Researchers, agencies and local planners are working towards improving our ability to understand, predict and deal with the impacts natural events have on communities and the environment [2]. This information is constantly being incorporated into local planning. And yet, hard lessons still to be learned by so much of the populace include doing one's homework, knowing what questions need to be answered, enlisting the services or qualified experienced professionals, and following their advice. Residents need to be prepared to deal with circumstances and emergencies that may pose a risk to both life and property. They need to site and build with a long-term perspective, and plan for the not-so-unexpected events that may cause damage or do harm. These principles are emphasized in the DVD, which we have found to have high impact, especially in-group settings, due to the power of the visual images coupled with the on-site role of coastal experts providing their context.

3. THE DEMO

The demo consists of snippets of the DVD "Living on the Edge: Building and Buying Property on the Oregon Coast," portions of which are being converted to mpeg or QuickTime for incorporation into the OCA itself (particularly "Learn" section, and "Hazards" therein). Excerpts will show how coastal storms and other natural processes pose particular challenges to those planning to build or buy property along the Oregon coast, and how viewers can learn more through exploration of the associated data sets and maps on the OCA. The full DVD is a co-production of Oregon Sea Grant at OSU and the state of Oregon OCMP. It also received its TV broadcast premiere on the Oregon Public Affairs Network (OPAN; www.opan.org) Thursday, Oct. 6, 2005.

4. ACKNOWLEDGMENTS

We wish to acknowledge NSF Digital Government grant #011359 as well as additional support from the NOAA Coastal Services Center and Oregon Sea Grant.

5. REFERENCES

- [1] Haddad, T., Haddad, T.C., Wright, D.J., Dailey, M., Klarin, P., Marra, J., Dana, R., and Revell, D. The tools of the Oregon Coastal Atlas. In Wright, D.J. and Scholz, A.J. (Eds.), *Place Matters: Geospatial Tools for Marine Science, Conservation and Management in the Pacific Northwest*, Corvallis, OR: Oregon State University Press, 134-151, 2005.
- [2] Komar, P.D., Diaz-Mendez, G., and Marra, J.J. Stability of the New River Spit, and the position of Oregon's beach-zone line, *Journal of Coastal Research*, 17(3): 625-635, 2001.
- [3] Ruggiero, P., Komar, P.D., McDougal, W.G., Marra, J.J. and Beach, R.A. Wave runup, extreme water levels and the erosion of properties backing beaches, *Journal of Coastal Research*, 17(2): 407-419, 2001.
- [4] Smith, C.L. Institutional mapping of Oregon coastal watershed management options, *Ocean & Coastal Management*, 45, 357-375, 2002.
- [5] Wright, D., Haddad, T., Klarin, P., Dana, R., and Dailey, M. Infrastructure for data sharing, spatial analysis, resource decision-making, and societal impact: The Oregon Coastal Atlas. In *Proceedings of The 5th Annual International Conference on Digital Government Research* (Seattle, WA, May 24-26, 2004), 131-132.

Integrating Information Technology and Social Science Research for Effective Government

MOST Policy Research Tool

Vincent Maugis

Management of Social Transformations Programme, UNESCO

SHS/SRP/POC

1 rue Miollis, 75015 Paris, France

+33-145-684-511

v.maugis@unesco.org

ABSTRACT

MOST Policy Research Tool will provide speedy access to policy-relevant comparative information. Users will be able to create individual research profiles based on subject categories, produce customized reports with select content from the original documents and easily build customized bibliographies. Innovative functions will also assist users to compare cases and assess the relevance of the policy options available.

SYSTEM DESCRIPTION

The Management of Social Transformations Programme (MOST) at UNESCO is initiating a no fee online policy research service. This project is expected to foster new modes of decision making based on actual evidence from situations in the community. As a matter of fact, the increasing need for relevant knowledge to inform international and national decision-making has overtaken current capacities of access, retrieval, organization and interpretation. This is a particularly acute problem for the social sciences. If social science is to be useful to policy makers, it needs to be accessible for comparison and verification. Without ready access to quality research conclusions, timely, focused and effective policy responses will be severely impeded at both national and international levels.

The free service will share results of research and policy recommendations on crucial human challenges worldwide. The resource has three elements:

- An international standard for policy documents, to allow their content to be held in a database;
- A huge and growing library of policy and research documents;
- An innovative web-based search tool.

The standard format for the documents makes that each element (each section in the format) acts as a knowledge item that is extractable and then becomes comparable across all documents. The tool's operational base is thus the resulting network of knowledge items.

The tool will deliver user-tailored, issue and location-specific policy-relevant material through a specially designed search function. Its focus will be on enabling easy access to high quality comparative social science research for decision-making. It will be developed in multiple languages, starting with English, French and Spanish with a view to expanding to the rest of the United Nations working languages. This service shall enable policies to be the "best possible" of options: evidence-based and linked to location-specific dynamics (context-sensitive) and also documented with assessments of similar experiences (best-informed). The primary objective is ensure actions will be better tailored to suit the specific needs of the populations concerned, which shall in turn experience improved living conditions.

How MOST makes a Difference – You are the Minister of Health:
In your country there are a million orphans whose parents have died from HIV/AIDS. You need policies for placing children in a host family, for their education, for the legal status of guardians... These policies must take into account the numbers of potential host families available and their resources, the capacity of the schools and the impacts on the social and economic life of the community.

With MOST Policy Research Service your staff enters the search terms into the free online server, and access the latest research on childcare, the policies of several other countries and the national social and economic statistics that will help you decide. With MOST your policy is built on evidence, on shared experience, on truly global knowledge.

How MOST makes a Difference - You are the city's Chief Planning Officer:

The city is huge, a multi-million person sprawl in a fast-developing tropical country. The pressure on you is immense - from the media, the politicians and from property developers. Some of the developers have tried to win you over, to encourage you to make city plans that would benefit their business. These developers have already bought some of the city's political

leaders. You need to create a rock-solid urban zoning plan, a plan that is based on facts, not on the preferences of the speculators.

To do this, you need evidence, and the best source for that is MOST Policy Research Service. There you will find the university research on industries in the city, the comparison of zoning in various cities (including yours) and the zoning policy documents of several other fast-growing tropical cities. MOST Policy Research Service gives you the evidence you need to create a fair, transparent, unbiased plan for your city and its people.

The system was designed by the Management of Social Transformations (MOST) Programme at UNESCO in cooperation with the Global System for Sustainable Development (GSSD) at the Massachusetts Institute of Technology (MIT) and in consultation with various National Commissions for UNESCO and government officials.

This Policy Research Tool addresses the following aspects of digital government research (from dg.02006 themes): information technology tools for government planning; public participation in democratic processes; transparency and usability; universal access to information and services; digital libraries and knowledge management; government processes and decision-making; public policy issues and implications

CONTACT PERSON

Vincent Maugis

Management of Social Transformations Programme
Section for Policy and International Cooperation
Division of Social Science, Research and Policy
Sector for Social and Human Sciences
U.N.E.S.C.O.

UNESCO
SHS/SRP/POC
1 rue Miollis
75015 Paris
France

Tel : +33-145-684-511
Fax : +33 -145-685-728
v.maugis@unesco.org

Integration of GIS and Educational Achievement Data for Education Policy Analysis and Decision-making

Sean W. Mulvenon, Kening Wang,
Sarah McKenzie, Denise Airola, and Travis Anderson
National Office for Research, Measurement and Evaluation Systems
University of Arkansas, Fayetteville, AR
seanm@uark.edu

ABSTRACT

Effective exploration and analysis of spatially referenced educational achievement data can help educational stakeholders and policy analysts accelerate interpretation of datasets to gain valuable insights. This demonstration will present a system developed in the National Office for Research on Measurement and Evaluation Systems (NORMES) for supporting web-based interactive exploration of state-wide educational data. The statewide statistical summaries are presented in thematic maps generated by Geographic Information System (GIS). Visual analysis of the thematic maps can help educational policy analysts to "see" complex spatial relationships and to detect patterns easily. Linking among GIS maps allows educators to interactively "drill down" through a hierarchy of geopolitical levels and to visualize statistical graphics and tables which are dynamically generated using Hypertext Preprocessor (PHP) script language. Through building up the relationships between maps, graphs, and databases, this interactive system allows administrators to gain an enhanced understanding of what is occurring in schools, and allow the policy analysts to generate more accurate and effective policy decisions.

1. PROJECT BACKGROUND

The implementation of "No Child Left Behind" (NCLB) legislation in 2001 has resulted in an exponential increase in the amount of educational data being collected. Educators and policy analysts need to explore large and multivariate educational datasets statewide to detect spatial or spatial-temporal trends in the datasets, and to examine the complex relationships between geographic aspects and other variables of interest. However, this presents special challenges for presenting the data in a readily understandable format with forms that are intelligible and insightful.

The National Office for Research on Measurement and Evaluation Systems (NORMES) has been creating and maintaining educational achievement data for the Arkansas De-

partment of Education (ADE) for six years. The nationally award-winning NORMES website (<http://normes.uark.edu>) has been designed specifically for dissemination of educational achievement data of Arkansas public schools. In order to support visual delivery and analysis of the spatially referenced educational achievement data for revealing spatial-temporal trends to educational policy analysts, Geographic Information Systems (GIS) was employed in the research. This demonstration will present a system which consists of two subsystems: Mapping Academic Performance in Schools (MAPS), which aims to support web-based interactive exploration of state-wide educational achievement data; Geographic Academic Policy Series (GAPS), which aims to provide a visual "snapshot" of achievement relative to current policy issues.

2. MAPPING ACADEMIC PERFORMANCE IN SCHOOLS

Three main components are essential to develop a web-based interactive data exploration system: a well-maintained website, expert created database and object-oriented thematic map series. The NORMES web site (<http://normes.uark.edu>) has been designed and maintained for dissemination of educational achievement data of Arkansas public schools. Development of state-wide academic achievement databases has also been completed in NORMES, and thus, is a readily available resource for use in this research project. The state-wide educational achievement databases include 1136 public schools, 254 school districts, which are distributed in 75 counties.

We have developed a demo system which can be located on our web site at <http://130.184.43.9/presentation>. We will use schools of the Fort Smith School District (in Sebastian County) to illustrate our approach. In this demo system, linking among GIS maps allows the educational stakeholders to "drill down" through geopolitical hierarchy levels, which consist of counties within the Arkansas State, school districts within counties, and public schools within school districts. The "drill down" ability will allow the educational stakeholders not only to know the spatial locations on GIS maps corresponding to the objectives of interest, but also will help the educational stakeholders search interested areas on the GIS map quickly and provide rapid access to substantial graphics and tables. The graphics and tables were linked to the thematic map dynamically. The thematic maps allow users to explore the spatial pattern; the inter-

actively displaying longitudinal statistical graphics let users view and compare the results among different years; and the detailed tables below the graphics allow users to explore the datasets in greater depth with the possibility of generating novel hypotheses.

3. GEOGRAPHIC ACADEMIC POLICY SERIES

The GAPS series displays maps in conjunction with state-wide educational statistical summaries, but does not require in depth understanding of statistics or methodology by the user. The goal of GAPS is to provide education policy analysis through geographical representations by using educational information and colors to represent the results for the analyses. Currently, GAPS examines the relationships between school district academic performance and other district level variables including percent of students participating in Free and Reduced Lunch Programs (a proxy measure for poverty), school district size, and per pupil spending. In addition, district performance on the ACT exam, which is completed by students intending to attend college, is presented and examined in relation to district size. GAPS also includes state maps identifying the academic performance status of all districts and schools in Arkansas. All of the research results are available on the website <http://normes.uark.edu/gaps>. Policy makers have been particularly interested in GAPS, noting that it provides them with an effective method for examining academic achievement statewide.

4. ACKNOWLEDGMENTS

We gratefully thank the Arkansas Department of Education for kindly providing us finance support and the data used in this study.

A Process-Driven Tool to Support Online Dispute Resolution

Lori Clarke, Alan Gaitenby, Daniel Gyllstrom,
Ethan Katsh, Matthew Marzilli,
Leon J. Osterweil, Norman K. Sondheimer,
Leah Wing, Alexander Wise
University of Massachusetts Amherst
Amherst, MA 01003
1-413-545-4228

LJO@cs.umass.edu

Daniel Rainey

National Mediation Board
1301 K Street, N.W., Suite 250 East
Washington, D.C. 20005
1-202-692-5051

Rainey@nmb.gov

ABSTRACT

This demonstration shows a prototype tool that projects an impression of how execution of a formally defined process will facilitate dispute resolution. Tool flexibility supports projecting the look and feel of a range of different processes, facilitating user evaluation of alternatives.

Categories and Subject Descriptors

K.4.3 [Organizational Impacts]: Requirements elicitation, prototyping, process definition

General Terms

Experimentation, Process, Dispute Resolution, Prototyping

Keywords

Online Dispute Resolution, Process Technology, Participatory Design, Grievance Mediation

1. INTRODUCTION

As the size and complexity of modern society continue to grow, the potential for disputes among the various parties in society grow as well. Indeed, the novelty of internet-based interactions is creating new opportunities for disputes. All of this creates the need for new and more efficient approaches to the burgeoning number and variety of disputes. Fortunately, technology also seems to offer approaches to their more efficient resolution. The field of Online Dispute Resolution (ODR) is exploring ways in which computer and communication technologies can facilitate dispute resolution while also decreasing the degree of involvement of humans [1]. ODR has been rapidly accepted in the commercial sector. But in government, where the need for increased dispute resolution efficiency is no less, acceptance has, nevertheless, been slower. Our project is exploring the premise that ODR acceptance in government can be expedited by facilitating the active involvement of diverse stakeholder groups

in the consideration and evaluation of dispute resolution approaches. We suggest, further, that this can be done by involving these stakeholders in the active consideration of various ODR approaches.

Our project is a collaboration of the University of Massachusetts with the National Mediation Board (NMB), the U.S. government agency charged with resolution of all labor-management disputes in the U.S. transportation industries (principally airlines and railroads). NMB has been seeing a steady increase in the need to mediate disputes, without commensurate increases in human resources. They have been interested in incorporating ODR into their work, but their continued credibility as an honest broker requires that they actively involve all of their many and diverse stakeholders in consideration of how ODR might be incorporated into their work.

2. OUR APPROACH

We view dispute resolution as a process, and have hypothesized that process definition, analysis, and execution technologies can be used both to provide automated support for dispute resolution and to effectively engage diverse stakeholder communities in consideration of just how this automated support is to be exploited. In particular, we regard ODR as a family of processes for dispute resolution in which computer and communications capabilities function as active agents in the conduct of the process. Clearly there are many possible ways in which dispute resolution processes might be defined, and many ways in which computer and communication technologies might function as active agents in each. This suggests that there is a need for research aimed at determining which processes are most appropriate under which circumstances, and which technologies are to be incorporated in which ways for best effect.

The goal of our research project is to demonstrate that process technologies can be effective in supporting this research. Our initial research has focused on using our Little-JIL process definition language to define precisely the IBB processes that NMB has been using [2]. One goal of this work is to prepare the way for our process execution capabilities to be used to marshal computer and communications capabilities to facilitate NMB's work. But, first we propose to use these process technologies to involve the various NMB stakeholders in active consideration of the processes that should be used. We believe that if all stakeholders have been actively involved in defining a dispute

resolution process, then they are more likely to be receptive to the acceptance of the outcome of a dispute resolution, even if that outcome seems unfavorable. We suggest that it is key that stakeholders have the ability to help define the dispute resolution process, and that they have the ability to monitor the execution of the process, to be sure that it conforms to the agreed-upon definition.

In the early phase of our project, we worked with NMB to use Little-JIL to define NMB's IBB process. Having done this, we began to suggest ways in which computer and communications technologies might be used as agents. It became clear that it was highly desirable to be able to project to the various stakeholders a concrete sense of the eventual look-and-feel of these various processes and their automation approaches. In order to support this capability we created a rapid prototype tool, called STORM, which creates a wide range of user interfaces to a correspondingly wide range of possible ODR systems. Our plan is to use STORM to acquaint the NMB stakeholders with the operational characteristics of various ODR approaches in order to gain their active and effective involvement in defining NMB's ODR strategy and technology adoption approach.

3. THE STORM PROTOTYPE

We found that NMB quickly became quite adept in understanding out Little-JIL process definition language sufficiently to be active and effective participants in defining the NMB IBB process. While this was gratifying, it became increasingly clear that the larger stakeholder communities were unlikely to be sufficiently adept in understanding the process definitions to be effective in critiquing them and debating the merits of various variants of the process and various automation approaches. Active engagement of all seemed to require that human stakeholders in the NMB processes would need to interact with an actual computer capability. Thus, we elected to create a suite of user interface capabilities in the early stages of our project in order to engage these stakeholders. This suite of user interfaces, coupled with a simplified backend data repository, comprises STORM, our prototype dispute resolution support system. Ultimately we will use our Juliette system that interprets Little-JIL processes to provide automatic presentation of specific dispute resolution processes to these stakeholder groups. These processes will enforce various constraints and disciplines, as mandated by the various processes. As STORM is merely a user interface suite, it will be unable to provide this enforcement. Humans will have to provide these constraints and enforcement to users of the prototype. But, insofar as such disciplined application of STORM

is provided, stakeholder groups will be able to evaluate various ODR approaches.

STORM uses the Tapestry toolset as the basis for its user interface capabilities. Tapestry provides a comprehensive suite of facilities for the creation of web-based applications, and offers considerable flexibility. Thus, STORM was constructed by an undergraduate student in a period of a few months. The stakeholder response to the STORM prototype was been uniformly positive and quite enthusiastic. We expect that STORM will indeed be an effective tool for involving diverse stakeholders in evaluation of various approaches to ODR in NMB.

4. NEXT STEPS

We are currently analyzing the initial responses to STORM. Initial results from our evaluation are sketched in a companion note [3]. We are evaluating the trade offs of expanding the functionality of the system. At the same time, we are exploring ways to realize the promise of the computer as a Fourth Party to fully "assist in identifying and evaluating interests, options and solutions".

5. ACKNOWLEDGMENTS

This material is based upon work supported by funds from the National Science Foundation under Grant No. IIS-0429297, as well as, funds from the National Mediation Board. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of the National Science Foundation or the National Mediation Board

6. REFERENCES

- [1] Katsh, E. and Rifkin, J., *Online Dispute Resolution: Resolving Disputes in Cyberspace*. San Francisco: Jossey-Bass, 2001.
- [2] Cass, A.G., Lerner, B.S., McCall, E.K., Osterweil, L.J., Sutton Jr., S.M., and Wise, A. *Little-JIL/Juliette: A Process Definition Language and Interpreter*. In *International Conference on Software Engineering*. Limerick, Ireland, 2000.
- [3] Katsh, E., Osterweil, L.J., Rainey, D., and Sonheimer, N.K. *Experimental Application of Process Technology to the Creation and Adoption of Online Dispute Resolution*. In *DG.o2006: The 7th Annual International Conference on Digital Government Research*. San Diego, 2006.

Supporting Humanitarian Relief Logistics Operations through Online Geocollaborative Knowledge Management

Brian M. Tomaszewski, Alan M. MacEachren, Scott Pezanowski, Xiaoyan Liu, and Ian Turton

Department of Geography and GeoVISTA Center

The Pennsylvania State University

University Park, PA 16802

(1+) 814-865-4448

{bmt139, maceachren, spezanowski, xiaoyan, ijt1} @psu.edu

ABSTRACT

Over the past two years, horrific disasters such as the Asian Tsunami, Hurricane Katrina, and the Pakistan Earthquake have demonstrated the critical need for effective technological infrastructure that is scientifically grounded in geo-visual group interaction theory [1] and humanitarian knowledge management procedures [2] to quickly and effectively facilitate planning for predictable events and post-event response. In this demonstration, we address specific issues that negatively impact the effectiveness of geocollaborative process in disaster relief. These include lack of common group operating picture, lack of command structure understanding and blatant miscommunication and misunderstanding about where relief supplies needed to be delivered, who will deliver them, when they need to be delivered, and the relevancy of deliveries to stricken areas. Our approach improves on existing systems by using methods and technologies that meet the challenges of coordinating the efforts of diverse and spatially distributed private, public, and governmental agencies throughout the world responding to disasters. This is accomplished by applying new forms of distributed geospatial data, technology, and collaboration functionality. We present our progress on the development of the Geocollaborative Web Portal (GWP), an asynchronous, open source geospatial information framework designed to support international group interaction and knowledge management in the context of humanitarian relief logistics.

Categories and Subject Descriptors

H.5.3 [Information Interfaces and Presentation]: Group and Organization Interfaces – *Asynchronous interaction, Collaborative computing, Computer-supported cooperative work, Web-based interaction.*

General Terms

Management, Design, Experimentation, Human Factors

Keywords

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Geocollaboration, Asynchronous Group Work, Logistics, Humanitarian Relief, Open Source GIS, Web Portal, Dynamic Web Map Services (WMS)/Web Feature Services (WFS) Integration, Concept Maps

1. INTRODUCTION

The Geocollaborative Web Portal (GWP) application (Figure 1) is designed to provide a common and intuitive interface through which asynchronous, geocollaborative activities can be conducted in support of humanitarian relief logistics operations. The GWP extends the core goals of the GeoCollaborative Crisis Management (GCCM) project [3, this volume]. It does so by providing specific functionality and tools within a web-based environment that support situation assessment, positioning and monitoring of field-teams and distribution sites, and supply routing. Special emphasis is placed on supporting international group interaction through collaborative annotation and visualization procedures, support for awareness of group interactions, multi-lingual map feature labeling, and organization-specific symbol sets to overcome communication barriers. In addition to facilitating asynchronous group interaction, the GWP enhances group knowledge development through the ability to integrate external WMS and WFS geospatial resources into the portal, access and author concept maps that represent operational rules and command structures in intuitive ways, store and retrieve file-based data resources such as site-imagery and documents, and monitor real-time RSS and GeoRSS feeds of situation-relevant information such as news and weather reports. While there has been independent research on most of these capabilities in other contexts, they have not (to our knowledge) been integrated previously within web-map / web-feature services, nor are they present in existing disaster systems.

2. ONLINE GEOCOLLABORATION

In our demonstration, we will focus on three components of online, map-based collaboration. The first of these is the concept of a *map session* where multiple users interact via a common web map space over an extended period of time. The map within the GWP uses the open-source MapBuilder API and JetSpeed Portal engine. These open frameworks have allowed our research team to develop scalable functionality and interface elements that can easily accommodate dynamic collaborative processes such as quickly adding or removing collaborators, tools and functionality as the situation dictates. Single user and subsequent group interaction is managed and persisted by GWP functionality. As

we will demonstrate, users have the option to be online at the same time interacting in near real time, and they can leave and return to sessions as needed, interacting asynchronously. GWP functionality tracks and records user map interactions such as panning, zooming, map extent, and annotation, and allows users to see what map actions others users have done, and where they have gone in both map space real world position (the latter through display of GPS tracks). This provides users of the application group with perceptual anchoring of actions taken by other users. Second, we will demonstrate the suite of tools that are available to individual users in the GWP. These tools allow users to input a diverse range of geospatial data into the portal and subsequently share it with other collaborators using a variety of methods. Capabilities include real-time address geocoding, GPX point and track parsing and rendering, geospatial image overlays, and dynamic WMS/WFS data source integration. In this portion of the demonstration, we will show how our underlying visualization and annotation procedures supplement data additions for an international audience through multi-lingual map feature labeling, and organization-specific symbol sets, and work toward overcoming issues of communication and understanding. Third, we will demonstrate the GWP's ability to integrate work with concept maps as well as with geographic maps. We are using concept maps to help collaborators structure knowledge about relief logistics procedures, understand how responsibilities and procedures for different organizations in a relief effort relate (or should relate), and identify tools and data relevant to specific situations.

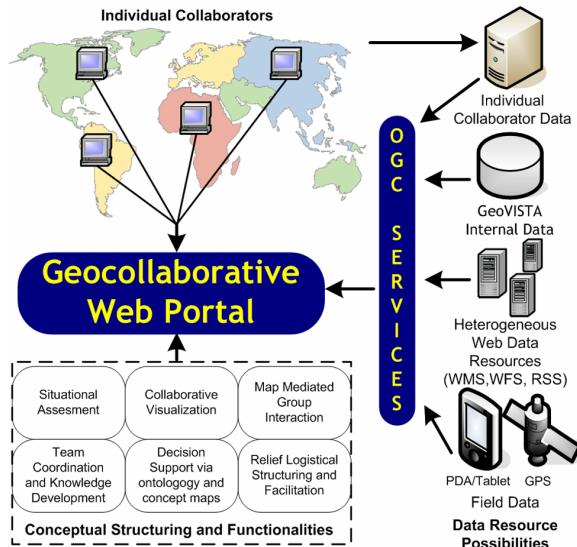


Figure 1. GWP Architecture Overview

3. APPLICATION SCENARIO

We demonstrate the utility of the GWP using a crisis scenario that illustrates how groups can collaborate to facilitate humanitarian relief logistics operations. This scenario demonstrates the myriad of geospatial, temporal, conceptual, and contextual issues and variables involved with international collaboration for disaster response, and how the GWP can overcome them. In order to determine the effectiveness of the GWP in relation to user interaction performance, we integrated the GWP with the

NeoCITIES [4] emergency management simulation application to develop group bench-mark performance measures.

Sample Relief Logistics Scenario:

A massive Tsunami strikes a heavily populated, island city located in the Pacific rim. The city had 6 hours to prepare for the impact. The broad effects of this include the disabling of military, government and civilian institutions and subsequent effects on transportation and communication structures. Civilians are evacuated to different parts of the island, and main sea and air ports of entry to the island are damaged, along with the island's main highway. As the situation unfolds, the GWP is used as a central online location for information sharing and reporting by all parties involved in responding to the situation, and to coordinate the development of the response logistics supply chain. GWP functionality also allows for temporally and spatially tracking situational urgency at various locations (lack of food, medical supplies), coordinating efficient and relevant resource allocation, enabling relocation of groups participating in the response as the situation develops, and identifying issues that arise from having a hierarchy of groups interacting with differing institutional perspectives (for example, a government agency won't share information with an NGO). Concept maps, organization-specific symbol sets, shared map annotation, external information feeds, and multi-language tools facilitate this collaboration in an international arena.

This scenario demonstrates how geocollaborative technologies, coupled with effective, intuitive information sharing can bridge potential language and cultural constraints between team members and lead to coordinated perspectives through the construction of team knowledge that can overcome issues inherent in disaster response collaboration.

4. ACKNOWLEDGMENTS

This work is supported by NSF Digital Government Research program under grant no. NSF-EIA-0306845. We acknowledge contributions from other team members, namely Marc Friedenberg, Joaquin Obieta and Adrian Cox.

5. REFERENCES

- [1] MacEachren, A. M. Moving geovisualization toward support for group work. In J. Dykes & A. M. MacEachren & M.-J. Kraak (Eds.), *Exploring Geovisualization*. Amsterdam: Elsevier. (2005). pp. 445-461
- [2] King, Dennis. Humanitarian Knowledge Management. *Proceedings of the Second International ISCRAM Conference*. Brussels, Belgium. (2005). pp. 1-6
- [3] MacEachren, A. M., McNeese, M., Cai, G., Fuhrmann, S., & Sharma, R. in press, Project Highlight: GeoCollaborative Crisis Management. *7th Annual National Conference on Digital Government Research*, San Diego, CA, May 21-24, (2006).
- [4] Michael D. McNeese, Priya Bains, Isaac Brewer, Cliff Brown, Erik S. Connors, Tyrone Jefferson, Jr., Rashaad E.T. Jones, and Ivanna Terrell. The Neocities Simulation: Understanding The Design And Experimental Methodology Used To Develop A Team Emergency Management Simulation. *49th Human Factors and Ergonomics Society Conference*, Orlando. (2005).

Opus (the Open Platform for Urban Simulation) and UrbanSim 4

Paul Waddell, Alan Borning, Hana Ševčíková, and David Socha

Center for Urban Simulation and Policy Analysis

University of Washington

Box 353055

Seattle, WA, 98195 USA

+1 206 221 4161

pwaddell@u.washington.edu, borning@cs.washington.edu,
hana@stat.washington.edu, socha@cs.washington.edu

ABSTRACT

This demo will give an introduction to Opus, the Open Platform for Urban Simulation, an Open Source platform for building simulations of land use, activity-based travel demand, and dynamic traffic assignment. It is a result of an international collaboration of research teams working on integrated land use, transportation and environmental modeling. We have developed a new version of UrbanSim – a simulation system for modeling urban development, originally demonstrated at the Digital Government 2004 Conference – as a component of Opus. We will demonstrate usage of UrbanSim for different stakeholder types, from modelers to policy makers.

Categories and Subject Descriptors

I.6.7 [Simulation and Modeling]: Simulation Support Systems.

I.6.3 [Simulation and Modeling]: Applications – *land use and travel modeling*. K.4.m [Computers and Society]: Miscellaneous – *urban planning*.

General Terms

Experimentation.

Keywords

Urban planning, modeling systems.

1. INTRODUCTION

Opus, the Open Platform for Urban Simulation, is a recent international collaboration to develop a robust, modular and extensible open source framework for land use, transportation and environmental modeling. It is an initiative to put model systems of different areas under one roof, and thus support their integration, increase their quality, facilitate increased collaboration among developers and users in the evolution of the

platform and its applications, and reduce the cost of building new model systems by leveraging a common framework [4].

Opus consists of independent packages, each of which usually represents a model system. Opus packages can, and usually do, use functionality of other Opus packages.

One of the main Opus packages is the urbansim package implementing the set of UrbanSim land-use models. UrbanSim (www.urbansim.org) is an open-source software-based simulation model for integrated planning and analysis of urban development, incorporating the interactions between land use, transportation, and public policy [2]. It is intended for use by Metropolitan Planning Organizations and others needing to interface existing travel models with new land use forecasting and analysis capabilities, and is planned to be the operational model for the Puget Sound Regional Council's four county area.

2. URBANSIM

Our previous version of UrbanSim, implemented in Java, was demonstrated two years ago at Digital Government Conference 2004 [3]. Since then, we rebuilt the entire system in order to address some significant shortfalls, and created a modular and extensible Opus package.

Here are some notable aspects of UrbanSim 4, the current version of the system:

- It is flexible, allowing people to easily experiment with the code, and construct new models by composing different parts.
- It is extensible. Adding a new model, variable, sampling method, estimator, data store, etc. can be done without touching the core code.
- It is scriptable, since it is written in Python.
- It is more accessible to modelers, since our experience so far is that they are much more open to using Python than Java.
- It is more transparent. For instance, the intermediate values are stored in a file system cache, which can later be mined to create indicator charts, maps, and tables (see Figure 1), or examined when debugging a problem.

- It has a fully integrated estimation process that shares code with the corresponding simulation procedure, and is almost completely automated (no more copying of results between an external estimation package and the simulation system). This eliminates a large and bothersome set of errors. It also dramatically reduces the time to estimate a model (hours instead of days). Re-estimating a model with different data takes just minutes or seconds.
- It has satisfactory performance, as a result of making extensive use of optimized C++ array and matrix manipulation libraries that are called from Python. For instance, as of this writing a 30-year simulation of the 16 UrbanSim land-use models over the Puget Sound Regional Council's dataset of 1.3 million households takes about 2 days (1 hour 40 minutes per year).
- It uses almost identical data as the prior version of UrbanSim, so only minor changes in the data structure are needed to convert UrbanSim 3.0 databases to work with the new UrbanSim.

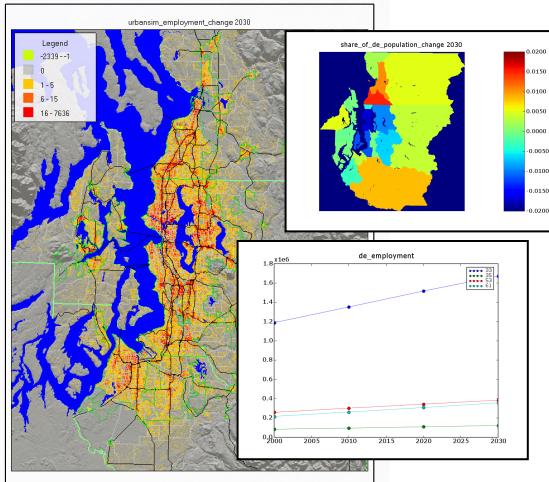


Figure 1 - Example indicators from UrbanSim.

3. OPUS COLLABORATIONS

The Opus version of UrbanSim is being transitioned into operational use for regional planning in our own region via our collaboration with Puget Sound Regional Council. As of this writing, several other Municipal Planning Organizations also are converting their data for the use of Opus.

As stated in the previous sections, Opus is designed as a platform that incorporates different kinds of simulation models from the land use, transportation and environmental area. For instance, we are collaborating with Professor Eric Miller's group at University of Toronto as they re-implement their Integrated Land Use Transportation Environment (ILUTE) modeling system to be used under the Opus umbrella. Professor Marina Alberti's group at the University of Washington has moved their land cover change model into Opus, and is planning on building additional environmental models in Opus. Furthermore, we are

collaborating with Michele Bierlaire of the École Polytechnique Fédérale de Lausanne and the Massachusetts Institute of Technology to create an Opus interface to the Biogeme package, which is an object-oriented software package designed for the maximum likelihood estimation of Generalized Extreme Value (GEV) models [1]. Several other groups expressed their interest in a close collaboration.

4. THE DEMO

Our demo will demonstrate all of the aspects listed above. It will start with simple experimental datasets, to illustrate the facilities that Opus provides for manipulating and visualizing data and running models. Then it will move to datasets from the full Puget Sound Regional Council application, which consists of over a million of households, demonstrating the ease of use of UrbanSim models on such complex data. It will illustrate how simulation and estimation work seamlessly together as two sides of the same model.

In addition, the demonstration will present a way of using the models by stakeholders not trained in any programming languages. We show that the same Opus models or model systems can be used interactively (for experimental purposes) as well as via a GUI (in a production mode).

Finally, we will show how indicator charts and maps can be generated using a variety of packages, including R, matplotlib, and OpenEV (see Figure 1).

5. ACKNOWLEDGMENTS

This research has been funded in part by the National Science Foundation Digital Government Program under Grant Nos. EIA-0121326 and IIS-0534094, and in part by a partnership with Puget Sound Regional Council.

6. REFERENCES

- [1] Bierlaire, M. BIOGEME: a free package for the estimation of discrete choice models. In *Proceedings of the 3rd Swiss Transportation Research Conference* (Ascona, Switzerland, 2003).
- [2] Waddell, P. UrbanSim: Modeling Urban Development for Land Use, Transportation and Environmental Planning. *Journal of the American Planning Association*, Vol. 68 No. 3, Summer 2002, pages 297-314.
- [3] Waddell, P. and Borning, A. A Case Study in Digital Government: Developing and Applying UrbanSim, a System for Simulating Urban Land Use, Transportation, and Environmental Impacts. *Social Science Computer Review*, Vol. 22 No. 1, February 2004, pages 37-51.
- [4] Waddell, P., Ševčíková, H., Socha, D., Miller, E. and Nagel, K. Opus: An Open Platform for Urban Simulation. Presented at the Computers in Urban Planning and Urban Management Conference, London, U.K., June, 2005,

POSTERS

Constituent-centric Municipal Government Coalition Portal

Nabil R. Adam, Vijay Atluri

Rutgers University
Newark, NJ 07102

adam@adam.rutgers.edu

atluri@cimic.rutgers.edu

Soon Ae Chun

City University of New York
Staten Island, NY 10314

chun@mail.csi.cuny.edu

Francisco Artigas, Irfan Bora,

Bob Ceberio

New Jersey Meadowlands Commission

Lyndhurst, New Jersey 07071

{francisco.artigas, irfan.bora,
bob.ceberio}@njmeadowlands.gov

1. INTRODUCTION

Businesses have used Customer Relationship Management (CRM) to gain strategic advantages by understanding and satisfying customer needs and creating short and long-term values. Similarly, governments face increased demands from citizens for better citizen-oriented one-stop services and more efficient and responsive government. Municipal governments started to introduce CRM to address these demands [2]. The key functionalities of CRM include (1) Citizen Segmentation into groups, such as individual taxpayers, corporate taxpayers, at-risk non-compliers; (2) Assess and monitor activities of certain segments of the population; (3) Provide tools, such as surveys, to uncover special needs and to measure service satisfaction levels; (4) Develop targeted offerings and outreach campaigns to meet specific population segment requirements, including education and notification of changes in tax laws.

The New Jersey Meadowlands Commission (NJMC) is a State agency created in 1969 with planning and permitting authority over an area containing 14 Municipalities in two counties known as the Meadowlands District. NJMC plays several roles, among them, as guardian for the preservation of the environmental and as facilitator for regional development and collaboration [3]. To promote the regional cost-saving operations among municipal governments and to help them leap-frog into electronic government practices, NJMC has initiated the “Municipal Technology Initiative” project (MTI). Its goals are to identify strategic IT areas for regionalizing e-government efforts through municipal government coalitions that will collectively save costs, benefit in service provisioning and enhance productivities.

A technology survey of municipal governments conducted during 2004-2005 to identify the target area for coalition-based E-government efforts showed that the Meadowlands municipalities are in different phases of e-government advances, ranging from catalog stage with mostly manual backend operations to well-advanced technological solutions in some specific areas [1]. Their technology solutions can be characterized by lack of planning and ad-hoc decision-making in purchasing, shortage of designated staff for hardware and software maintenance and updates, and great need for computer skills training of

government employees. They exhibited great similarity in organizational structure, business operations and types of services. Also, in common is that they face great demands for better services by constituents, a great burden of compliance reporting to the State government, and a great need to collaborate among municipalities in terms of sharing resources. It became clear from the survey that municipalities could greatly benefit and lower their costs by forming municipal coalitions and collaborating to address their common constituent-related needs and services, with a CRM system which would otherwise be cost prohibitive to undertake separately.

In this paper, we present the current on-going project that provides an *infrastructure solution* called Constituent-centric Municipal Coalition Portal (CIMCOP) for the Meadowlands District's municipal governments. The infrastructure aims to provide customized services using Constituent Relationship Management for their respective constituents, including citizens, businesses, and other governments. Its architecture, design principles, benefits and challenges are described.

2. CONSTITUENT-CENTRIC MUNICIPAL GOVERNMENT COALITION PORTAL

The goal of this project is to provide an information and service infrastructure networking together fourteen municipal governments in the NJ Meadowlands area using Constituent Relationship Management for their primary constituents, including citizens, businesses and other governments. Its specific goal is to allow all fourteen municipalities be more responsive to their constituents, and to allow the citizens to be more participatory, not just consuming municipal services and information, but also voicing their concerns for the towns and Meadowlands in general, ultimately shaping the policies and enhancing government accountability. In addition, all towns, not just technologically advanced towns, will be bootstrapped to provide more advanced citizen services through this portal solution without duplicating costs and efforts. This solution also aims to foster the collaboration among the fourteen governments in terms of purchasing, training and emergency responses, and meet the municipal government's compliance requirements to the State and federal governments.

Design Principles: The development utilizes the following principles for success: (1) the municipal governments are autonomous entities. The existing services and operational systems are not replaced but leveraged and information and data are seamlessly funneled to them from the CIMCOP. (2) it aims to maximize the commonalities (services and forms) among municipalities; (3) the services provided are functionally complete; (4) it provides universal access through an intuitive and

easy interface for any types of constituents and devices; (5) it maximizes interoperability in data sharing and collaboration; (6) the constituents' sensitive data is protected through proper security measures.

Architecture: The Constituent Relationship Management for the municipal government coalition requires an infrastructure of networked distributed municipality systems and services, as shown in Figure 1.

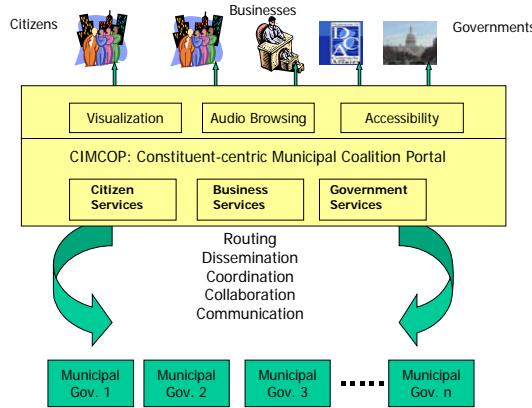


Figure 1 Municipal Government Coalition Portal

Specific functionalities for the coalition CRM portal include:

- To provide common citizen services (online utility payments, paying parking violation, volunteer service enrollment, map services) and to provide common compliance services (building permits, vehicle permits, subdivision applications, zoning map services)
- To support information search and download (employment opportunities, grants and contracts for businesses, election information, environmental protection activities and publications, library locations) and to support automated distribution of citizen data, forms, documents to individual municipal governments and to appropriate officials (e.g. citizen requests for a street repair, a dog license form, State budget report forms)
- To foster citizen participation in decision making (town's or council meeting information (agenda, minutes), an electronic submission of their voices before and after the meetings) and to understand citizens' needs and perceptions through feedback and log analysis
- To support the collaboration and coalition among municipal governments (collective purchasing service, collective training course information, sharing of information regarding emergency responses and criminal investigations, sharing of equipment.)

Issues and Challenges: Customer Relationship Management in businesses focuses on processes of marketing, selling, order management and call/service center. The adoption of an existing CRM system may not be the appropriate solution for the municipal coalition. Various processes require regulatory knowledge. The customized contents and services may be challenging, given the lack of service history and citizen profiles for different constituents, due to mostly manual operations. The updates of the contents and backend operations should be managed automatically and accurately.

Approach: Our approach is to categorize the services and information common to all town services into a topic ontology. Each category can have services, information, and forms for each town. The Web contents can be automatically generated through the topic categories customized for a specific town with similar structural layout information. In addition, we utilize inter-agency workflow and Web services to integrate processes with the regulatory ontology. The portal employs access control mechanisms for citizens' sensitive information. To enhance accessibility, we employ voice-browsing technology for summarizing vast amounts of government information, and presenting them with audio output [4]. We utilize visualization using graphs and Google map Web services [5].

Current development focus areas for four constituents include: *Citizens* (request for service and information, service status tracking, smart form design, environmental data access.), *Businesses* (business registration, development permit application and tracking system, land development permit identification system, business location decision support system); *Governments* (online annual budget reporting system to the State; equipment sharing among municipalities; map sharing); *Government employees* (administration of service requests, event calendar management, environmental decision support system).

3. CONCLUSIONS

The coalition-based municipal Web-portal and infrastructure will provide municipal governments with the following benefits: (1) be more responsive to citizen needs, (2) provide more efficient government, (3) provide greater information access to decision makers, (4) enable better alignment between public policy and the needs of the public, (5) enable proactive management of "hot-button" issues, (6) provide 24 hour availability of access to information and feedback for constituents, and (7) help the public become better informed and more satisfied. Similarly, it will also benefit NJMC (1) to be more informed and more responsive to individual municipal governments, and their constituents' needs, (2) to have more comprehensive evaluations of Meadowlands-wide policies, (3) to leverage the service commonalities in municipal governments, (4) to enables Governments to aggregate and analyze citizen needs and perceptions in individual towns and the whole region, and (5) to foster communication and exchange of information for better regional planning.

4. ACKNOWLEDGMENTS

This work is supported by the Municipal Technology Initiative grant under the MOU between the New Jersey Meadowlands Commission and CIMIC, Rutgers University.

5. REFERENCES

- [1] Municipal Technology Initiative, Final Report, CIMIC-Rutgers University, CIMIC Technical Report July-01-2005, July 2005.
- [2] B. Larsen and M. Milakovich, Citizen Relationship Management and E-Government, Electronic Government, *Proceedings of 4th International Conference, EGOV 2005, LNCS 591*, Springer, 2005, pp 57-68.
- [3] New Jersey Meadowlands Commission, *A Meadowlands Renaissance, 2004/2005 Annual Report*, 2005.
- [4] Voice browser Activity: <http://www.w3.org/Voice/>
- [5] Google Maps API: <http://www.google.com/apis/maps/>

Semantics-based Threat Structure Mining

N. Adam¹, V. Atluri¹, V. P. Janeja¹, A. Paliwal¹, M. Youssef², S. Chun³,
J. Cooper⁴, J. Paczkowski⁴, C. Bornhoevd⁵, I. Nassi⁵, J. Schaper⁵

¹CIMIC, Rutgers University

²Arab Academy for Science and Technology

³City University of New York

⁴The Port Authority of New York and New Jersey

⁵SAP Labs

ABSTRACT

Today's National and Interstate border control agencies are flooded with alerts generated from various monitoring devices. There is an urgent need to uncover potential threats to effectively respond to an event. In this paper, we propose a *Semantic Threat Mining* approach, to discover threats using the spatio-temporal and semantic relationships among events and data. We represent the potentially dangerous collusion relationships with a *Semantic Graph*. Using domain-specific ontology of known dangerous relationships, we construct an *Enhanced Semantic Graph* (ESG) by scoring the edges of the semantic graph and prune it. We further analyze ESG using centrality, cliques and isomorphism to mine the threat patterns. We present a Semantic Threat Mining prototype system in the domain of known dangerous combination of chemicals used in explosives.

1. INTRODUCTION

The Port Authority of New York/New Jersey (PA) manages and maintains bridges, tunnels, bus terminals, airports, PATH commuter trains, and the seaport around New York and New Jersey that are critical to the bi-state region's trade and transportation capabilities. The continuous monitoring of cars, trucks, trains, and passengers is a necessary precaution for preventing major threats to safety. The amount of data and potential alerts generated from these monitoring activities are enormous and heterogeneous in nature due to the different types of monitoring devices, ranging from text messages to images, audios and video feeds. The challenge is to mine and identify meaningful potential threats, and minimize false alerts. Important is an ability to infer threats coming from several independent seemingly benign activities. Often ignored is the threats implicated when these independent activities are looked at together as illustrated in the following scenario.

Motivating Example: Consider a customs office inspecting a truck shipment carrying liquid Urea entering through the port in Los Angeles, whose final destination is Phoenix, AZ. Assume there

is another shipment entering through the port in Newark carrying cyclotrimethylene trinitramine (RDX) is bound for Wintersburg, AZ.

The two shipments, when viewed in isolation, appear to be benign. However, the spatial proximity (shipments with spatially close destinations), temporal proximity (the two events occurring close in time), and semantic proximity (the materials being shipped have some semantic relationship, for example, can be combined to make explosives), would indicate possible collusions among entities and enhance a potential threat discovery and detection capability. Here it is essential to look at spatio-temporal proximities first since purely semantic proximity may lead to frivolous and non-relevant threat structures to be identified.

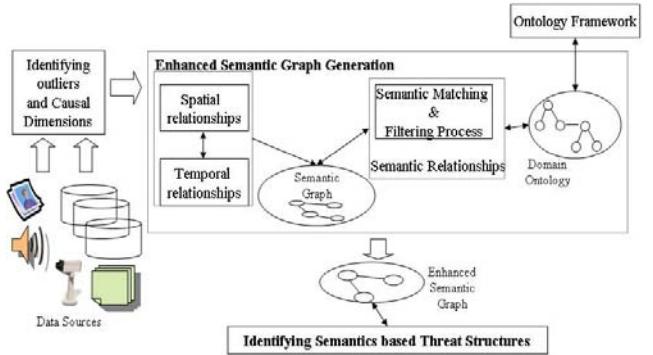


Figure 1 Semantic Threat Structure Mining Approach

2. SEMANTICS DRIVEN DATA MINING

Our approach depicted in Figure 1 consists of the following distinct steps: *i) Semantic Graph (SG) generation by outlier detection*: We use data mining to generate nodes in SG by identifying interesting entities namely outliers and their causal dimensions [3]. *ii) Enhanced Semantic Graph (ESG) Generation*: The connectivity between the nodes of SG are established and pruned to generate an enhanced Semantic graph (ESG) by the following two-steps. a) First we identify *spatio temporal relationships* between the outliers; b) Second, we identify *semantic relationships* and *semantic scores* between the outliers, using domain ontologies and reasoning. The semantic enhancement includes removing relations that are not supported by the reasoning using the semantic relationship scores between the dimensions. *iii) Identification of Threat Structures*: The ESG is further analyzed for the semantic centrality, semantic cliques and isomorphic paths to identify semantics based threat structures.

3. SDM Prototype System

3.1 Dataset

We have tested our approach on the PIERS data, comprising of imports, exports data and U.S. and overseas profiles of companies. The PIERS data comes from multiple ports and agencies. Moreover, some shipments, when observed closely, may lead to suspicious terrorist behavior, which is analogous to the threat structures considered in this paper.

For the prototype system, we have labeled outliers by visual inspection. We have vertically partitioned it to allocate a set of dimensions to each domain. The domain experts from the Foreign Operations Division, U.S. Department of Homeland Security have been identifying the semantic relationships within the dimensions of the PIERS data and labeling of the outliers in this data. However for checking the accuracy and efficacy of outlier detection we discuss detailed results with other datasets [3].

3.2 Semantic Matching for Relationship Mining

To identify the semantically dangerous relationships among outlier data sets, we used a domain specific ontology on Threat agents; specifically chemical threats constructed using Protégé as shown in Figure 2. It includes two major taxonomies: the domain specific concept taxonomy, e.g., types of chemicals, and an operational taxonomy for the matching process that we refer to as the *Potentially Dangerous Combinations* (PDC).

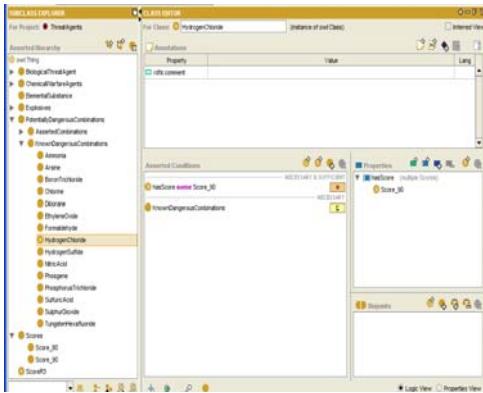


Figure 2 Threat Agents Ontology

The data set is matched against the ontologies and reasoned using the RACER reasoner [2]. The matching mechanism first reads the ontology from an OWL file, converts it to DIG format [1], second receives a potentially dangerous combination of chemicals and creates a parent concept as the union of this combination, and third performs the matching and returns the enhanced semantic graph.

3.3 Identification of threat structures

Once we identify the semantic relationships generating the ESG, we utilize it to identify threat structures based on the top weighted semantic relationships, other semantic properties such as semantic cliques, semantic centrality, and semantic isomorphic paths.

We have vertically partitioned the PIERS data into 3 domains such that each domain consists of some part of each record. Interrelationships between outliers are labeled based on the outliers detected and the semantic weights generated by the ontology framework. It was observed that only 50% of the labeled interrelationships were identified. We believe that this loss is due to

the manual partitioning and labeling of the data where, some actual outliers may have been lost. In Figure 3, the weighted graph in the left of the window shows the semantic graph, the right top part of the window shows the Enhanced Semantic graph with the weights in terms of the semantic weights generated by the matching process. The bottom right part shows the discovered top interrelationships and the semantic centrality. Although we discover all interrelationships for presentation we show top 10 relationships.

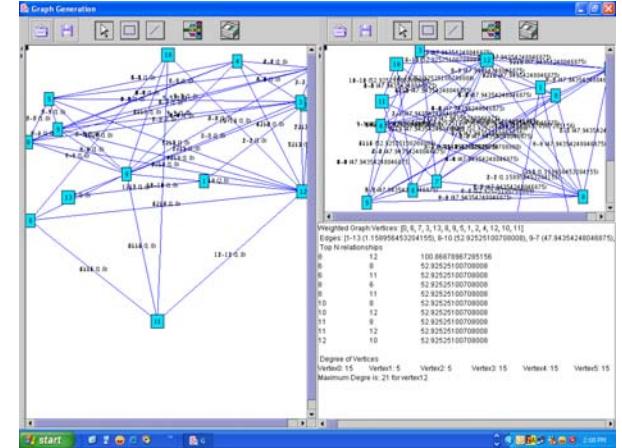


Figure 3 SG and ESG of PIERS data

4. CONCLUSION AND FUTURE WORK

In this paper, we presented an approach to enhance the semantic graphs by discovering collusion relationships existing in spatio-temporal and semantic dimensions among events, using the concept of the Enhanced Semantic Graphs (ESG). The potential threats are identified with semantic centrality, semantic cliques and isomorphic paths. As part of our future research, we propose to focus on the general problem of identifying relationships between normal entities without restricting ourselves to simply outliers. We plan to evaluate existing Semantic graphs generated from text data. Consistency across systems in determining semantic distances and the robustness of such calculations is essential in homeland security domain. We need to investigate determining the relative semantic distance between two concepts through an inspection of the values of selected attributes through a hierarchy of variables ranging from those that most directly related, termed proximate variables, to those most distantly removed, termed ultimate variables.

9. ACKNOWLEDGMENTS

This work is supported in part by the National Science Foundation under grant IIS-0306838 and SAP Labs, LLC.

10. REFERENCES

- [1] S. Bechhofer, "The DIG Description Logic Interface: DIG/1.1", Available from <http://dl-web.man.ac.uk/dig/2003/02/interface.pdf>, 2003.
- [2] V. Haarslev, R. Moller, "RACER User's Guide and Reference Manual", Available at <http://www.cse.concordia.ca/%7Ehaarslev/racer/racer-manual-1-7-19.pdf>, 2000.
- [3] V.P.Janeja, V.Alturi, J.S.Vaidya, and N.Adam. Collusion set detection through outlier discovery. In IEEE Intelligence and Security Informatics, 2005.

Automated Dental Identification System (ADIS)

Hany Ammar, Robert Howell

West Virginia University

ammar@csee.wvu.edu

rhowell@wvu.edu

Mohamed Abdel-Mottaleb

University of Miami

mottaleb@miami.edu

Anil Jain

Michigan State University

jain@cse.msu.edu

Law enforcement agencies have exploited biometrics for decades as key tools in forensic identification. With the evolution in information technology and the huge volume of cases that need to be investigated by forensic specialists, automating the process of forensic identification became inevitable. Postmortem (PM) identification, encountered in mass disasters (e.g. wars, natural disasters, etc), requires the use of biometric characteristics that resist early decay of body tissues as well as withstand severe environmental conditions. To this end, dental features are the best candidates for PM identification [1].

In 1997, the Criminal Justice Information Services (CJIS) division of the FBI created a Dental Task Force (DTF) to foster the creation of an Automated Dental Identification System (ADIS). ADIS will provide automated search and matching capabilities for digitized radiographs and photographic images, so as to generate a short match list for dental forensic experts.

Research teams from West Virginia University (WVU), Michigan State University (MSU), and University of Miami (UM), in coordination with CJIS, are collaboratively developing a research prototype of ADIS. Creating ADIS requires the development of a highly automated environment that integrates image processing and pattern recognition techniques thus achieving both high accuracy and quick response time. To this end, we are not only automating feature extraction and matching, but we are also analyzing the radiographs in order to utilize underlying image structures that are often difficult to assess merely by visual examination [2].

Proposed ADIS Architecture

ADIS consists of the following main components (see Figure 1): (i) Record Pre-processing component, (ii) Potential Matches Search component, and (iii) Image Comparison component. The record pre-processing component handles the following tasks: (a) record cropping into dental films [3], (b) enhancement of films to compensate for possible poor contrast, (c) classification of films into bitewing, periapical, or panoramic views [4], (d) segmentation of teeth from films [5], and (e) annotating teeth with labels corresponding to their location [6]. The potential matches search component manages archiving and retrieval of dental records based on high-level dental features (e.g. number of teeth and their shape properties) and produces a candidate list. Several approaches to realize potential matches are presented in [7][8]. The image comparison component conducts low-level tooth-to-tooth comparison between subject teeth -after alignment [9]- and the corresponding teeth of each candidate, producing a short match list [10]. The philosophy behind the architecture of the prototype ADIS is to exploit high-level features for fast retrieval of a candidate list produced by the

potential matches search component and then to refine the candidate list using low-level image features, leading to a short match list.

Web-ADIS Environment is built on a 3-tier architecture consisting of database layer, client layer and server layer. Database layer is used to store the dental record along with the description data. Client layer takes care of the presentation of the details to clients and the business logic and transaction management is taken care of by the server layer. Web-ADIS Client is a user using a Java enabled browser which sends requests to ADIS Server which processes the requests with the help of database server to send back the results to the client. The webADIS client is technically a Java-enabled web browser that connects to the webADIS Server. The webADIS Server commands the sequence of events in the application. The immediate subcomponents of the webADIS Server are the Web Server, the ADIS server, the Pre-processing Server, the Potential Match Server, and Image Comparison Server. Web Server securely connects the Web-ADIS Server to the outside world through the Internet. ADIS Server acts as a controller for webADIS and controls the flow of objects to the subsequent servers and to the Database Server. In Database Server the dental/non-dental features of dental records are stored in a Feature database. The corresponding images of the dental record are stored in the image database also called the Digital Image Repository (DIR) [11].

Web-ADIS has three modes of operation: configuration mode, identification mode and maintenance mode. Configuration mode helps the user to tune the system to use realizations of his choice and test the system by applying the configuration on the system. Identification Mode can be used by the client to get the matches for the submitted record of the subject and the Maintenance mode is meant to upload the database sever with new reference records and also to update the preprocessing server, potential match server and image comparison server with new realizations that are likely to give better results [11].

Preliminary Testing of the Image comparison

The image comparison component conducts low-level tooth-to-tooth comparison between subject teeth, after alignment [12], and the corresponding teeth of each candidate, thus producing a short match list [10]. Because dental radiographic films capture projections of distinct teeth; and often multiple views for each of the distinct teeth, Nassar [13] looks for a scheme that exploits teeth multiplicity to achieve more reliable match decisions when we compare the dental records of a subject and a candidate match. He proposes a hierarchical fusion scheme that utilizes both aspects of teeth multiplicity for improving teeth-level (micro) and case-level (macro) decision-making. The achieved genuine accept rate is approximately 85%.

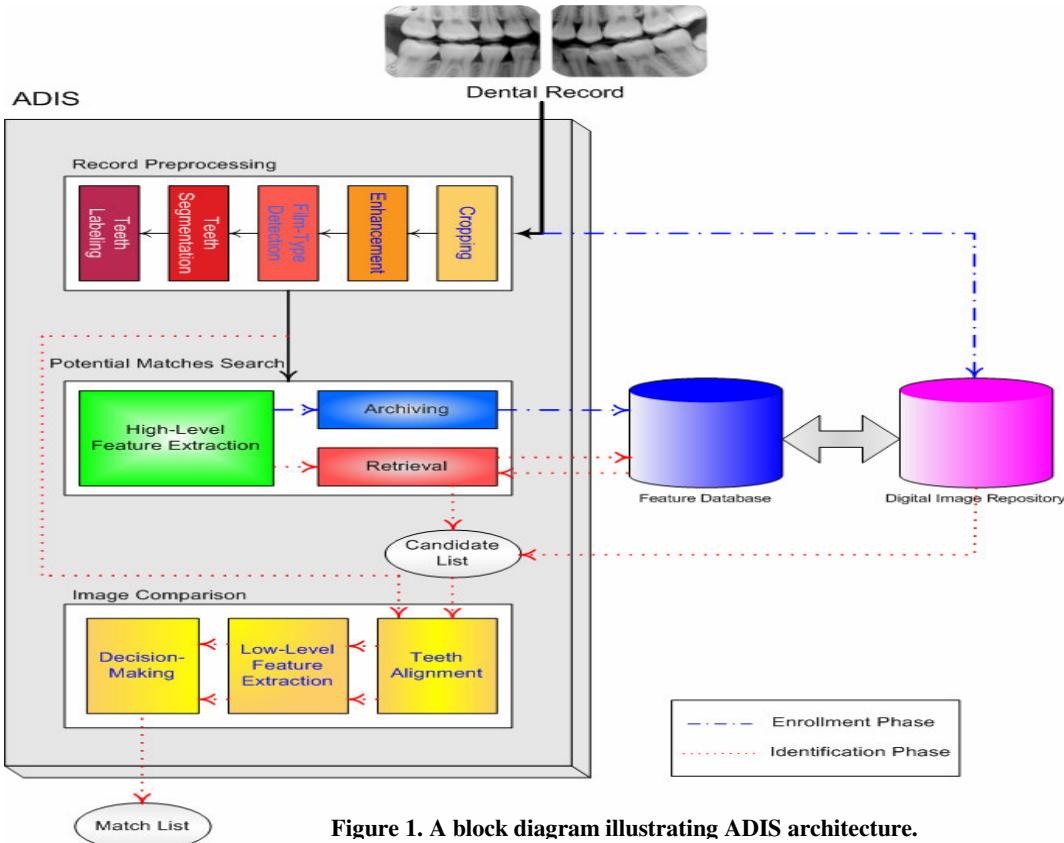


Figure 1. A block diagram illustrating ADIS architecture.

REFERENCES

- [1] American Society of Forensic Odontology, *Forensic Odontology News*, Vol. 16, No. 2, 1997.
- [2] Fahmy, G., Nassar, D., Haj Said, E., Ammar, H., Abdel-Mottaleb, M., Chen, H., Jain, A., "Toward an Automated Dental Identification System", *Journal of Electronic Imaging*, Vol. 14, No. 4, 2005.
- [3] Li, X., Abaza, A., Nassar, D., and Ammar, H. "Fast and Accurate Segmentation of Dental X-ray Records", *Proc. of 2006 International Conference on Biometric Authentication (ICBA)*, Hong Kong, Jan 2006.
- [4] Zhou, J., and Abdel-Mottaleb, M. "A Content-based System for Human Identification based on Bitewing Dental X-Ray Images", *Pattern Recognition*, Vol. 38, No. 11, pp. 2132-2142, 2005.
- [5] Haj-Said, E., Nassar, D., Fahmy, G., and Ammar, H. "Dental X-ray Image Segmentation", *Proc. of SPIE Technologies for Homeland Security and Law Enforcement conference*, Orlando, FL, 12-16 April 2004.
- [6] Mahoor, M., and Abdel-Mottaleb, M. "Automatic Classification of Teeth in Bitewing Images", *Proc. of ICIP 2004*, Singapore, August 2004.
- [7] H. Chen and A. Jain, "Dental Biometrics: Alignment and Matching of Dental Radiographs", *IEEE Transactions on PAMI*, Vol. 27, No. 8, pp. 1319-1326, August 2005.
- [8] Nomair, O., and Abdel-Mottaleb, M. "A system for human identification from X-ray dental radiographs", *Pattern Recognition*, Vol. 38, pp. 1295-1305, 2005.
- [9] Ogirala, M. "Multi Resolution Dental Image Registration based on Genetic Algorithm" Masters Thesis, Lane Department of Computer Science and Electrical Engineering, West Virginia University, May 2005.
- [10] Ammar, H., and Nassar, D. "A Neural Network System for Dental Radiograph Comparison", *Proc. of 2nd IEEE International Symposium on Signal Processing and Information Technology*, Marrakesh - Morocco, Dec 2002.
- [11] Chekuri, S., Nassar, D., Abaza, A., Bahu, A., Ammar, H. and Fahmy, G., "A web-based Automated Dental Identification System (webADIS)". *Proc. of 5th IBIMA International Conference on Internet & Information Technology in Modern Organizations*, Dec 2005.
- [12] Nassar, D., Ogirala, M., Adjerooh, D., and Ammar, H., "An Efficient Multi-Resolution GA approach to Dental Image Alignment", *Proc. of SPIE Electronic Imaging*, Jan 2006.
- [13] Nassar, D., "Automated Dental Identification: A Micro-Macro Decision-Making Approach", Ph.D. Dissertation, Lane Department of Computer Science and Electrical Engineering, West Virginia University, Dec 2005.

An Empirical Study on E-Government Readiness: The Roles of Institutional Efficiency and Interpersonal Trust

Kallol Bagchi

College of Business,

University of Texas at El Paso,

Email: kbagchi@utep.edu

Stuart Gallup

College of Business,

Florida Atlantic University,

Robert Cerveny

College of Business,

Florida Atlantic University

Email: sgallup@fau.edu Email: cerveny@fau.edu

ABSTRACT

The use of information technology and communications to advance the interaction between a government and its citizens is expanding at an increasing rate. Governments around the world are looking for avenues to exploit interactive tools to enhance public awareness and reduce operating costs. A primary obstacle to successful implementation of e-government is trust. In this preliminary study, we investigated the dimensions of institutional efficiency and trust and their effect on the deployment of e-government and found preliminary evidence that these indicators play a significant role in e-government readiness.

1. INTRODUCTION

The topic of E-government is receiving much attention among IS researchers [8, 9, 13, 10]. Areas of research interest include the transformation of government and democratic processes, each requiring an interactive relationship between government and its citizens [14]. E-government is defined as „a (scheme) to improve the relationship between the private citizen and the public sector through enhanced, cost-effective and efficient delivery of services, information and knowledge. It is the practical realization of the best that government has to offer” [17].

E-government programs can provide citizens with information faster and more efficiently, such as on-line filling of tax forms. Other benefits include: improved government management, a decrease in corruption, revenue growth, and/or cost reductions [1]. The Cape Gemini report indicates several examples from Europe of e-government in operation such as:

- Denmark’s electronic invoicing saves €150 million annually in the administrations
- In Romania electronic procurement has reduced procurement costs by almost 25%
- In the Netherlands 1/3 of all students are using an online student-grant service which is visited by 70,000 people each month.
- Disabled people now get immediate benefits in Belgium, where previously it took 3-4 weeks and significant paper handling.

E-Government Readiness is measured as “how willing and ready the governments around the world are to employ the opportunities offered by Information and Communications Technology (ICT) to improve the access, and quality, of basic social services to the people for sustainable human development” [17]. The top 10 ranked nations in 2004 based on the E-government readiness index is presented in Table 1. The United States is at the top of the list, closely followed by Western European nations and two Asian nations. As expected, all of these are developed nations. In fact, over 90% of public service providers in Europe now have an on-line presence, and 40% of basic public services are fully interactive [1]. This paper contributes to the body of literature by providing empirical support to the claim that trust and institutional efficiency matter in E-government readiness. We also

show that these two dimensions along with government presence and sophistication of government on-line services contribute to e-government readiness.

Table 1. The Top 10 2003 Nations

E-government readiness Index: Top 10 countries			
<i>United States</i>	0.9132	<i>Australia</i>	0.8377
<i>Denmark</i>	0.9047	<i>Canada</i>	0.8369
<i>United Kingdom</i>	0.8852	<i>Singapore</i>	0.834
<i>Sweden</i>	0.8741	<i>Finland</i>	0.8239
<i>Republic of Korea</i>	0.8575	<i>Norway</i>	0.8178

2. THE CONCEPTUAL MODEL

There are many potential factors (economic, institutional, social etc.) that have not been fully investigated that could contribute to e-government readiness. For example, trust and institutional efficiency are two such factors. The conceptual model examined in this paper is shown in figure 1 and is described next.

2.1 Trust

Trust is known for reducing transaction costs in economic activities because less time is needed to investigate business partners [3, 20]. Trust takes on a more critical role in a global setting because trust varies substantially across countries [20]. Knack and Keefer [12] found that higher levels of trust are instrumental for growth in 29 market economies. Indeed, trust is a crucial enabler in e-commerce, where consumers are exposed to far lesser degrees of dependence and risk, affecting purchase intentions, inquiry intentions and sharing personal information [4, 7, 11].

For citizens to develop trust in the agency managing the online tax process, it is advisable that an independent third party be engaged to build and certify that the system will behave in a certain manner. The agency must also demonstrate that the system is managed by trustworthy people [18]. This implies the efficient institutional practices (such as an efficient judicial system for fair and efficient disposal of disputes, better rule of law, less corruption etc) play a positive role in increased e-government related transactions.

One of the measures of e-government readiness is the web presence of various branches of the government. A higher degree of presence can mean that these nations are in advanced stages of e-government development.

We next describe the measures used to determine web presence versus e-government sophistication. The degree to which government agencies use on-line services can determine the sophistication of their e-government environment. Services go through phases of sophistication, from an “information-only”

website to a full transaction and case-handling capability. Note that the sophistication of services, here measured as GOVSERV03, is not the same as the web presence measured as GOVPR03; at best, there exists a weak relationship between the two ($R^2=0.19$).

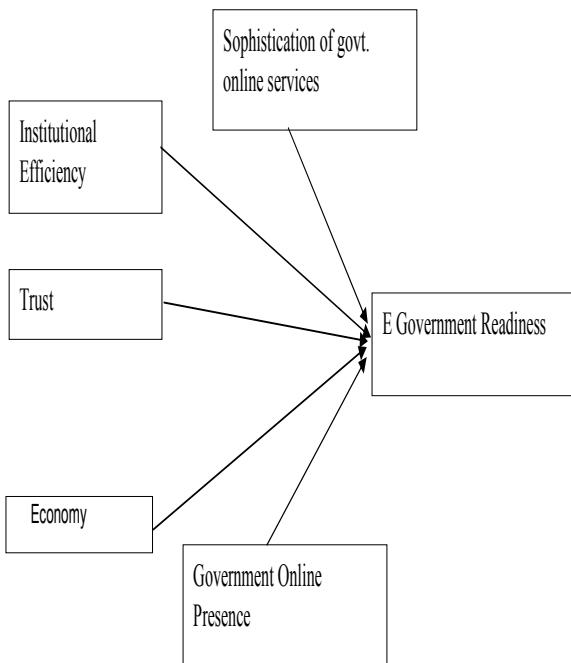


Figure 1. The Conceptual Model

Examples of sophisticated e-government environments could be electronic tax-submission system, interactive forms, and system that enable feedback from citizens on various social issues

Our control variable is the economic status of a nation, captured by GDP. It has been shown empirically that Information Technology (IT) and human development are two factors that impact e-government development [13]. Both of these are strongly related to GDP; the human development index has economy as a component.

3. DATA

The nations considered in this study are: *India, Nigeria, China, Romania, Turkey, Poland, Bulgaria, Chile, Czech, S. Africa, Lithuania, Hungary, Argentina, Brazil, Mexico, Belarus, Russia, Latvia, Estonia, Portugal, S. Korea, Ireland, Slovenia, Spain, The U.K., Italy, Netherlands, Belgium, Austria, France, Canada, The U.S., Iceland, Germany, Denmark, Finland, Norway, Sweden, Japan and Switzerland*.

A total of forty nations are considered due to data availability and cross-tabulating data from other sources. The data sources used were databases of the World Bank, United Nations, Heritage Foundation, World Value Survey (WVS) and World Economic

Forum (WEF). The United Nation (UN) Global E-Government Survey 2003 [16] presents a comparative ranking of countries in the world.

The E-Government Readiness Index is a composite index comprising:

1. Web Measure Index
2. Telecommunication Infrastructure Index
3. Human Capital Index

For example, of the set of 191 nations studied, the U.S has a value of 0.92 (the highest), followed by Sweden (0.84), Australia (0.83) and Denmark (0.82).

4. RESULTS AND DISCUSSIONS

We ran an OLS regression with all the major relevant indicators. The results are shown in Table 2. Multicollinearity was not significant in the regression as the range of Variance Inflation Factor (VIF) values were within the range 1.6-3.7.

Table 2. The E-Government Readiness Regression

PREDICTOR VARIABLES	Dep. Variable EGOV03				
	COEFFICIENT	STD ERROR	STUDENT'S T	P	VIF
CONSTANT	-0.32946	0.22293	-1.48	0.1489	
LNGDP	0.08381	0.01956	4.28	0.0001	3.7
TRUST	0.00124	6.70E-04	1.85	0.0734	1.6
GOV ONLINE PRESENCE, 2003	0.02661	0.01085	2.45	0.0197	2.3
HERITAGE INSTITUTIONAL INDEX 2003	-0.04307	0.02309	-1.87	0.071	3.5
SOPHISTICATION OF GOVERNMENT ONLINE SERVICES, 2003	0.0238	0.01038	2.29	0.0284	2.0

Adjusted R-Square 0.886; N=38, (F =60.43, p< 0.0000)

The results of this preliminary study suggest that e-government readiness is affected by interpersonal trust, institutional efficiency, government online presence and sophistication of government online services, after controlling for GDP. This study suggests that interpersonal trust and institutional efficiency are factors that contribute to the success of an e-government initiative. Future research is needed to further explore the dimensions of interpersonal trust and institutional efficiency together with many other cultural, social and institutional variables that can affect an e-government system.

5. REFERENCES

Available upon request

The BioPortal Project: A National Center of Excellence for Infectious Disease Informatics

Hsinchun Chen

Daniel Zeng

Chunju Tseng

Cathy Larson

Management Information Systems Department, The University of Arizona

McClelland Hall 430, 1130 East Helen Street

Tucson, AZ 85721

{hchen, zeng, cal}@eller.arizona.edu; and chunju@email.arizona.edu

ABSTRACT

In this project summary, we briefly present the technical objectives and accomplishments of our Infectious Disease Informatics project. We describe the inter-agency, inter-disciplinary, and academia-government partnerships critical to project success and discuss the broader application context for this research.

Categories and Subject Descriptors

H.4.2 [Information Systems]: Types of Systems – *decision support*

General Terms

Algorithms, Human Factors

Keywords

infectious disease informatics; technology adoption; data analysis; visualization; hotspot analysis; biosurveillance

1. INTRODUCTION

Funded by the National Science Foundation Digital Government program through grants #EIA-9983304 and #ITR-0428241, the National Center of Excellence for Infectious Disease Informatics project aims at goals of: a) developing an integrated and scalable information sharing, monitoring and analysis environment across jurisdictions and species for major infectious diseases; b) developing novel data analysis, surveillance, and visualization techniques to meet the critical needs of infectious disease informatics (IDI) for human, animal and plants; and c) gaining a systematic understanding of related policy, user evaluation and technology adoption issues.

This effort represents a critical initial step towards the full-scale implementation of a national infectious disease information infrastructure (NIDII). A prototype system called BioPortal has been developed, implementing many of the functionalities required for a NIDII. BioPortal is available for demonstration at <http://www.bioportal.org>. More information about BioPortal and related technical research is available in [1] and [2].

2. PROJECT ACCOMPLISHMENTS

Our project was initiated in October 2003 under the guidance of the Infectious Disease Informatics Working Committee, a federal inter-agency committee. The BioPortal system initially provided integrated, Web-enabled access to distributed data sources related to two prominent infectious diseases: West Nile Virus and Botulism. Protocols and agreements for adding additional data sources have since been well-established.

Additional animal and public health information has been incorporated into the BioPortal system, including international Foot-and-Mouth Disease (FMD) datasets and related gene sequence data, and emergency room chief complains (for syndromic surveillance purposes). The BioPortal's infectious disease data query and analysis functions are summarized below:

- **Data Access.** BioPortal provides customized query interfaces to the available infectious disease datasets. For almost all of these datasets, spatial and temporal coordinates of the disease cases and sightings/test results, among others, are essential. Both coordinates can be presented at different granularities (e.g., for location, specific street address/county/state; for time, specific day and time/weekday/week/month/year). Users often need to perform queries and then aggregate the findings based on various location/time granularity requirements. BioPortal provides a flexible tabular tool that allows users to select a preferred granularity level and presents the summary data accordingly.
- **Data Analysis.** BioPortal supports spatial-temporal clustering analysis, i.e., hotspot analysis, using various methods for detecting unusual spatial and temporal clusters of events. Such event clusters can facilitate disease outbreak detection and predictive modeling. As part of BioPortal research, we have developed a new hotspot analysis method based on support vector machines, which has been shown to outperform existing methods using both simulated and real-world datasets [1].
- **Data Visualization.** BioPortal includes Spatial-Temporal Visualizer (STV), a visualization tool which allows users to effectively explore spatial and temporal patterns, based on an integrated tool set consisting of a GIS tool, a timeline tool, and a periodic pattern tool. Recently a new component has been added to STV facilitating gene sequence analysis and comparison based on phylogenetic trees. This component extends the STV in an important dimension that is critically needed to analyze disease outbreaks involving multiple strains of a virus (e.g., in the case of FMD).

We also completed a preliminary user evaluation study to collect feedback on various aspects of the BioPortal implementation [3]. Initial findings indicate that both user information satisfaction and end-user satisfaction are significantly higher with BioPortal than with the benchmark program (the spreadsheets traditionally used for infectious disease data analysis). As a system, BioPortal was considered more usable than the benchmark program by our subjects, who also perceived BioPortal to be more useful and easier to use than the benchmark program. Follow-up evaluation studies are underway with a larger pool of subjects and infectious disease analysis tasks of greater variety.

3. PARTNERSHIPS

In the initial phase of the project targeting at developing an integrated West Nile Virus and Botulism data portal, our interdisciplinary research team included three groups: (1) the Artificial Intelligence Lab at the University of Arizona, (2) the New York State Department of Health and its partner Health Research, Inc., and (3) the California State Department of Health Services and its partner PHFE Management Solutions. The National Biological Information Infrastructure/National Wildlife Health Center, as part of U.S. Geological Surveys, has also been an active research partner. Our current research team has been expanded to include the University of Utah, the University of California Davis, the Kansas State University, and the Arizona State Department of Health Services (AZDHS). This team composition represents a balance of significant Information Technology research and system development experience, public health domain expertise, livestock health expertise, infectious disease data analysis experience, and user evaluation competence. In addition to data sharing, three collaborative projects with new partners with funding support from sources outside of the NSF have been developed to expand the initial BioPortal functionalities and dataset coverage, and to beta test and evaluate the BioPortal infrastructure and the NSF-funded research in real-world applications. We briefly summarize these derivative projects below.

- *International FMD Monitoring and Surveillance.* The FMD virus (FMDv) is considered to be one of the most contagious infectious disease agents of domestic animals. The estimated costs of an FMD outbreak to the food and agricultural sectors of the U.S. vary widely but are always counted in the tens of billions of dollars. Partnering with the FMD Lab, UC Davis, we aim to (a) establish sustainable cooperative agreements with countries, agencies, and laboratories for acquisition of FMDv isolates and epidemiological data, (b) provide in real time and to the public via the FMD BioPortal sequence and related epidemiological information about new and important FMDv isolates, and (c) develop new temporal-spatial-genetic models to predict phylogenetic changes in the virus.
- *Syndromic Surveillance.* Syndromic surveillance aims to provide an early estimation of the population's health status. Working with AZDHS, we are integrating several syndromic surveillance datasets (mostly on emergency room chief complaints) from the Phoenix area into BioPortal. A customized version of the BioPortal system tailored for syndromic surveillance is being developed and will be deployed at AZDHS.
- *Livestock Health Surveillance.* The ability to detect emerging animal disease outbreaks at early stages is critically needed as part of U.S. homeland security efforts. The Rapid Syndrome Validation Project – Animal (RSVP-A) system, developed by Kansas State University (KSU), is aimed to develop such a capability and has demonstrated that information technology can help detect various kinds of livestock anomalies by collecting and analyzing livestock syndromic data. We are currently working on integrating RSVP-A with

BioPortal to leverage RSVP-A's advanced real-time data collection capabilities and BioPortal's hotspot analysis and visualization functions. The long-term goal of this project is to develop a comprehensive set of tools for animal health data sharing, access, analysis, and surveillance.

4. APPLICATION CONTEXT AND PROJECT MANAGEMENT

The BioPortal project addresses the interdependencies among disparate and distributed disease data systems. Owing to the real-time sharing of data that BioPortal could provide, the system could benefit public health agencies in their infectious disease fighting activities (preventing, detecting, managing) across jurisdictions. We also see potential applications for law enforcement and national security concerning biological terror attacks, especially with regard to surveillance and hotspot analysis capabilities. The technological and policy issues explored in this project are also applicable to other digital government domains.

The University of Arizona team is the primary grantee with the other teams as subcontractors. Providing appropriate funding support to our data providers, domain experts, users and evaluators, is key to maintaining the partnership and promoting productive interactions. Extensive communications between team members, especially in this distributed, cross-jurisdictional context, are critical to the success of our project. We communicate through regularly scheduled conference calls, and collaborate on papers and presentations for a wide variety of professional, academic and technical audiences. Project objectives are mutually agreed upon, and system demonstrations are examined carefully by all participants. Such close working relationships are key to achieving success and sustaining a multi-year research project. At the same time, they enable us to pursue additional funding and project opportunities that aim to operationalize BioPortal-enabled applications and have a potentially significant real-world impact.

5. ACKNOWLEDGMENTS

Research supported in part by the National Science Foundation through Digital Government Grant #EIA-9983304 and Information Technology Research Grant #IIS-0428241.

6. REFERENCES

- [1] Zeng, D., Chang, W., and Chen, H. (2004). "A Comparative Study of Spatio-Temporal Data Analysis Techniques in Security Informatics," in Proceedings of the 7th IEEE International Conference on Intelligent Transportation Systems, pp. 106-111.
- [2] Zeng, D., Chen, H., Tseng, L., Larson, C., Eidson, M., Gotham, I., Lynch, C., and Ascher, M. (2004). "West Nile Virus and Botulism Portal: A Case Study in Infectious Disease Informatics," in Intelligence and Security Informatics, Proceedings of ISI-2004, Lecture Notes in Computer Science, pp. 28-41, Vol. 3073, Chen, H., et al. (eds.), Springer.
- [3] Hu, P., Zeng, D., Chen, H., Larson, C., Chang, W., and Tseng, C. (2005). "Evaluating an Infectious Disease Information Sharing and Analysis System," in Intelligence and Security Informatics, Proceedings of ISI-2005, Lecture Notes in Computer Science, Vol. 3495, Kantor, P., Muresan, G., Roberts, F., Zeng, D., Wang, F.-Y., Chen, H., and Merkle, R. (eds.), Springer.

Understanding the Adoption and Diffusion of Innovative Information Technology Curricula: A Case Application to Master of Public Administration Programs

Shu-Chuan Chiu

School of Public and Environmental Affairs
Indiana University Bloomington
1217 E. Maxwell Lane,
Bloomington, IN 47401

Email Address: shuchiu@indiana.edu

ABSTRACT

Information technology (IT) management training provided by Master of Public Policy or Administration (MPA) programs has evolved generally from software applications embedded in traditional public policy or management courses to dedicated courses on IT management or policy. However, the degree to which individual MPA programs have evolved varies. Literature in this field has focused on providing status reports on IT education in MPA programs and making curricular recommendations. Building on the foundation, this study tries to explain MPA programs' variation in IT management education by using institutional and resource characteristics of these programs. Data collected from 183 MPA programs in the U.S. show that program ranking and accreditation status appear to be significantly correlated to whether a MPA program has a concentration on IT management. This study will have implications for countries or localities that strive to improve public service by providing cutting-edge IT management training to future public managers. It will also help inform policy makers who would like to provide effective incentives to motivate programs to better IT management training.

General Terms

Management, Human Factors, and Theory.

Keywords

Diffusion of innovations, public administration education, e-government, information technology management, resource dependence theory, institutional theory, and curricular adaptation.

1. INTRODUCTION

The emergence of information technology (IT) as a policy issue in public administration (PA) education has reflected the growing importance of IT in the public sector. Over the last several decades, public sector investments in IT and the development of e-government initiatives have increased significantly. Indeed, nowadays IT is embedded in almost everything public organizations try to accomplish, and the manifestation of the influence of computing on government has evolved from automation, using computer applications, record keeping, searching, and information provision to two-way transactions and more sophisticated integration across different agencies, different levels of government, and different sectors. In addition, issues

related to IT, such as information network security, IT acquisition and policy, outsourcing, and the "digital divide," to name just a few, have been emerging. These issues require public managers to be knowledgeable in the area at the intersection of IT and its related ethical, financial, administrative, legal, regulatory, cultural, and social dimensions of IT management.

Regarding the proper venue for IT management training, it is sometimes suggested that the on-the-job training approach would best fit each organization's IT tasks. However, such training tends to be application-centered without sufficient appreciation of the aforementioned dimensions involved in using IT in the public sector; nor does it provide the requisite methodology to tackle complex IT management tasks. Therefore, some PA scholars have called for MPA programs to take on the challenge of strengthening future managers' IT management capacity.

Studies specifically dedicated to the IT in MPA education have generally focused on updating IT education status in MPA programs and making curricular recommendations. In these studies, little explanation has been given as to why MPA curricula show different degrees of deviation from these recommendations; nor has research in this area tried to link empirical studies to theory testing or building.

To begin to fill in this knowledge gap, this paper tries to increase the understanding about why IT management curricula differ among MPA programs. The next section discusses methodology and data sources, followed by analysis results. The final section of the paper offers concluding remarks.

2. Methods

To approach MPA programs' variation in their IT management curricula, this study uses a dependent variable that is operationalized as whether or not a MPA program offers IT management or a similar area as one of the concentrations or specializations. Since it is a dummy variable, the major analytical tool adopted is probit regression procedures. The explanatory variables are mainly related to the institutional and resource environments of these programs. Specifically, these variables include full-time faculty size, whether the MPA program is housed in a public administration department or school, total credit hours required for the degree, availability of other programs or departments on the same campus that offer IT management training, program ranking (from the U.S. News and World Reports, USNWR), and whether the program is accredited by the National Association of Schools of Public Affairs and

Administration (NASPAA). The selection of these variables is informed by institutional theory proposed by DiMaggio and Powell, Meyer and Rowan, and Tolbert and Zucker, as well as resource dependence theory by Pfeffer and Salancik.

The sample this study uses is derived from combining the following three sources of MPA programs lists: (1) the member institutions database from the Association for Public Policy Analysis and Management (APPAM) website, (2) the NASPAA roster of accredited programs, and (3) the list of the master of public affairs programs ranked by the USNWR. After removing overlapping programs, there are 183 MPA, Master of Public Policy, or similar programs in the sample. In this paper I use the term "MPA" to refer to all these programs. Data then were mostly gathered from the websites of these programs. In the cases where online information was incomplete, unclear, or cannot be found, I contacted the programs via email to clarify.

3. Results

Thanks to IT, data collection was implemented smoothly as all programs in the sample publish their program and curricular information on the World Wide Web, thereby greatly reduced the problem of missing data from low response rates that plagued some previous studies that relied on surveys to gather data. Due to the methodology of aforementioned sampling procedure, the sample is encompassing and representative of the academic programs at issue.

The data confirm the variation in IT management curricula as well as the resource environment among MPA programs. 30 out of 183 programs explicitly offer IT management or a similar area as a concentration, and 27 programs require course(s) on IT management. Titles of these required courses range from introductory courses to IT, Government Information Systems, to courses addressing the organizational and cultural implications of IT. Further, programs on average have 16 faculty members, but the range of faculty size is considerable – some programs have single-digit numbers of faculty members, while a few programs have more than 100 members on the faculty. Nevertheless, it is more uniform for MPA programs to have at least one other academic program on the same campus that offers IT management training; only 59 programs do not have such resources. As to program affiliation, 88 MPA programs are housed in a School of Public Affairs or similar schools whose titles are public policy or PA-centric, while others are affiliated to a college or school that consists of more diverse academic programs. 45 MPA programs are part of a Department of Public Administration or a similar department, as opposed to a Department of Political Science or the like.

Additionally, a probit model shows USNWR ranking and NASPAA accreditation status are the only two explanatory variables that show statistically significant relations to whether a MPA program has a concentration on IT management. Such results provide evidence to support the hypothesis derived from institutional theory, which would argue that the ranking of a MPA program reflects the program's overall conformity to the ranking institutions' criteria. Likewise, by setting curricular standards, including standards for IT components, NASPAA plays a role in influencing MPA programs' curricular adaptation. Even though NASPAA does not have a minimum course requirement for each curricular standard, at least some MPA programs appear to have

responded to these standards, and this study's results support the hypothesis that NASPAA-accredited programs are more likely to provide an IT management concentration than those that are not accredited. The results also support the hypothesis that the higher the ranking, the more likely the program offers an IT management concentration. From an institutional view, the more the program responds to such overall standards, the more likely it will adopt the components in the standards, as both are part of institutional rules programs strive to conform to.

The lack of statistical significant results for the other explanatory variables may be related to data quality. Since programs update the information on their websites at different rates, the accuracy of the data this study relies on vary. Further, the measurement of "faculty size" is constrained by inconsistent definitions of "faculty" across programs, and often there is not enough information on the Web to decipher whether a faculty listing includes only the core faculty of the program, all faculty affiliated to the department or school, or all of the above plus joint appointments. Moreover, the titles of the departments or schools that house the MPA programs do not necessarily mirror certain types of program inclination as some previous research suggests. For one thing, the titles might be results of historical legacies and the programs have evolved. For the other, the categorization of program affiliation is arguably arbitrary due to the inherent difficulty in having a clear-cut criterion to judge whether a program is affiliated to a PA department or school.

Given all the aforementioned constraints, this study can still offer policy implications for people who are interested in MPA education in general, and in particular for people who are concerned about the IT management education in MPA programs.

4. CONCLUSION

As IT has intertwined with an increasing scope of issues and tasks, managers' IT management capacity has become more consequential in the success of the overall performance and effectiveness of organizations. However, IT management training has seen varied degrees of diffusion into MPA curricula. This research attempts to account for this variation by linking varied IT management curricular development to MPA program characteristics based on perspectives of resource dependence theory and institutional theory.

The results of this research show that the gaps between MPA programs continue to exist, in terms of both resources and their IT curricula. In addition, the results also show that ranking and accreditation status predict reliably the IT management curricular development status.

These findings' major implications are two-fold. First, the finding about the influence of institutional rules from ranking and accreditation entities on innovative IT curricula can provide insights into the diffusion of innovations, which in turn can help facilitate the dissemination of changes to meet the challenges in various areas, such as technology infrastructure security, digital divide, interoperability, civic engagement, etc. Second, it will help bridge e-government and public administration education research by understanding the diffusion of IT management training that MPA programs offer to future public managers, whose leadership and IT capacity have been identified in the e-government literature to be critical for the success of e-government projects.

Quality Evaluation of e-Government Digital Services

Flavio Corradini, Francesco De Angelis, Alberto Polzonetti, Barbara Re

name.surname@unicam.it
Dip. Matematica e Informatica
Università di Camerino
Via Madonna delle Carceri, 9
Camerino (MC), Italy

ABSTRACT

In this paper we present a “quality estimation model” for digital e-Government services suitable for quality evaluation, monitoring, discovery, selection and composition.

Categories and Subject Descriptors

D.2.8 [Software Engineering]: Metrics; D.2.9 [Software Engineering]: Managements

Keywords

e-Government, Quality of Service, quality model

1. INTRODUCTION

The promotion of e-Government by Public Administrations (PA) encourages the development of digital services. To provide an approach to quantify Quality of Service in e-Government digital service we must take into consideration the following components: (a) study of parameters to be included in the model; (b) introduction of a mathematic model to define utility function; (c) representation of parameters and metrics using a shared representation.

In this paper we take in consideration the second step. We present a “quality evaluation model” for digital e-Government services. This is useful for services quality evaluation, monitoring, discovery[1], selection[3] and composition [4].

2. MODEL FOR QOS ESTIMATION

Quality model for QoS quantification of e-Government digital services take into consideration the work in [2]. This paper describes how to formally discriminate between n distinct Web Services in input and for all services they associate m evaluation parameters. For every service an overall quality value is defined from these inputs. Our study introduces further elements that aren't provided in literature but that must be considered in the e-Government domain. We study input data homogenization to handle inconsistency on metrics on e-Government parameters. Also, we must introduce

the interaction between them. It measures dynamic relation between service parameters in such way to indicate how a parameter behaviour condition each other.

Proposed model considers an unique service for evaluation based on quality parameters that represent it. Discovery concept in e-Government is still premature. Indeed, in this domain we don't speak about offer market but there is an unique offer to satisfy demands of specific users target. Every parameter is combined with a weight related to interaction level reached by parameter in respect to the other. QoS computation considers every parameter in a group that are able to associate similar criteria. It is more rational referring to cost than execution cost and transmission cost in a separate way. Every group is weighting related to group feature importance to final QoS value. To evaluate overall service quality is necessary to consider the steps in Figure 1 and experimental results from the model. This is exposed in the following subsections.

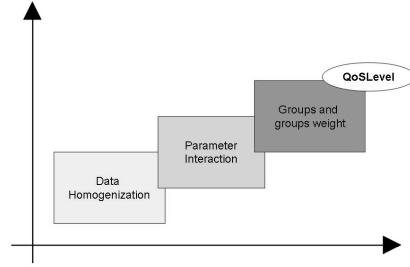


Figure 1: QoS evaluation model stack

2.1 Model Phases

Phase 1: Data Homogenization

At first we analyze data homogenization of array Q that contains n quality parameter observed during measurement process ($Q = \{q_1, q_2, \dots, q_n\}$). Through a normalization process we report parameter value on a increasing graduation when min is 0 and max is 100. For this purpose we foresee two arrays, N e C . Array N assumes discreet value 0 and 1 to differentiate proportional growing and inverse proportional growing of quality parameter in respect to overall quality. The second vector indicated by C represents max

value that q_i parameter in Q vector may assume. Value associated to this vector are binded to the study on single parameter and it depends by specific metrics used to express parameter other than methodology applied to measurement. For homogenization purpose we define function 1 such that for each element q_i in vector Q we obtain a new element in vector Q' .

$$f_1(q_i, n_i, c_i) = n_i \left(\frac{q_i * 100}{c_i} \right) + (1 - n_i) \left(100 - \frac{q_i * 100}{c_i} \right) \quad (1)$$

Phase 2: Parameter Interaction

In the second phase we consider interaction among different quality indicators. To this purpose we provide MI matrix that describes in detail relative interaction between parameters. This matrix consider interaction $\varepsilon(q_j, q_k)$ for every pair (q_j, q_k) of elements in vector Q . $\varepsilon(q_j, q_k)$ value is comprised between 0 and 1 and it measures interaction reported in MI according to (2).

$$m_{j,k} = \begin{cases} \varepsilon(q_j, q_k) & \text{se } \varepsilon(q_j, q_k) > 0 \\ 0 & \text{se } \varepsilon(q_j, q_k) \leq 0 \end{cases} \quad (2)$$

Clearly, $\varepsilon(q_j, q_k) > 0$ shows a positive interaction, while $\varepsilon(q_j, q_k) \leq 0$ shows absence of interaction. We have interaction matrix MI with diagonal value equal to 0 to represent insignificant interaction of parameter with itself.

Starting from MI matrix we define a new vector P that measures *interaction weight* between quality parameters. We suppose, for example, that cost interacts positively with the user's perceived trust, these two parameters may be associated to interaction factor that considers this observation. From this vector we extract *interaction factor* φ_j of parameter q_j .

$$\varphi_j = \frac{p_j}{n - 1} \quad (3)$$

In (3) we show the parameter interaction level respect to the max number of interactions and it doesn't consider recursive phenomenon on parameter legitimate by null value on matrix diagonal. Indeed, it is not admissible a situation when cost parameter interact with itself. Analysis between distinct services may be different because cost parameter of one service may interact with cost of other services. Considering interaction factor vector, every element of vector Q' must be normalized to obtain a new vector Q'' based on function in (4).

$$f_2(\varphi_i, q'_i) = \varphi_i * q'_i \quad (4)$$

Phase 3: Grouping and Group Weight

At this point we consider that in the model the parameters can be grouped and managed like group of features with different importance. For this reason we introduce matrix D and vector G . D shows parameter that can be considered together because it refers to similar features, while G represents element importance in respect to the others. Moreover, with l we refer to total number of groups of quality criteria. Applying matrix D to Q'' , with $D * Q''$ we obtain vector G . To this point it is necessary to know groups cardinality displayed by h_i .

In the next step we present an array F of weight associated

to every group. In this manner we can specify user preferences or developer preference over group i or over single parameter. In this case we have groups formed only by a single parameter.

Finally, it is possible to evaluate an overall quality value for a service considering QoS function showed in (5).

$$QoSLevel = \frac{\sum_{i=1}^l \frac{g_i}{h_i} * f_i}{\sum_{i=1}^l f_i} \quad (5)$$

3. CONCLUSION

QoS in e-Government domain covers a fundamental role. Either service provider or service user can differentiate, estimate and reuse services with the same ability. In this work we present a model for quality estimation and guideline to take into consideration quality aspects.

The most important experimental results show that $QoSLevel$ is a linear function, it increases or decreases steadily in respect of the parameters. From analysis of frequency distribution we can see that it follows a normal trend. Moreover, the approximation steadily improves as the number of observations increases. Finally, we notice that the upper bound for quality depends on the interactions between the parameters. If the interaction decreases the quality level has low values, while if the interaction and the interaction factor increase also the $QoSLevel$ increases. All this things are legitimated by the fact that frequency distribution, in case of considering minimal interactions, are attested near low values, while stretch to normal with the highest interaction. The increase of the interaction among parameters supports the quality of the proposed model because the e-Government process stimulates different depending factors.

Acknowledgements

We would like to thank Fausto Marcantoni for stimulating comments on parts of this paper.

4. REFERENCES

- [1] E. Merelli F. Corradini, C. Ercoli and B. Re. An agent-based matchmaker (A case study in biomedical services discovery). In *proceeding of WOA04 Sistemi complessi e agenti razionali*, pages 150–156, September 2004.
- [2] Y. L., A. H. Ngu, and L. Z. Zeng. QoS computation and policing in dynamic web service selection. In *WWW Alt. '04: Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*, pages 66–73, New York, NY, USA, 2004. ACM Press.
- [3] E. M.l Maximilien and M. P. Singh. A framework and ontology for dynamic web services selection. *IEEE Computer Society*, 4:84 – 92, September - October 2004.
- [4] Chris Pelz. Web Services Orchestration and Choreography. *Computer*, 36(10):46–52, 2003.

The Role of Public Return on Investment Assessment in Government IT Projects

Anthony M. Cresswell

Center for Technology in Government

187 Wolf Road, Suite 301

Albany, NY 12205

tcresswell@ctg.albany.edu

1. INTRODUCTION

This poster will present the preliminary results of a study of how public return on investment assessments and concepts are used in government IT investment decisions and developments. After decades of investments in information technology, running into billions of dollars, governments are largely unable to convincingly demonstrate a return on investment that is widely understood or based upon well-grounded measures. Nevertheless, it is clear that much in government has been dramatically changed by IT and many programs and services are believed to be more effective and less expensive as a result. Unfortunately, it remains difficult to confirm those beliefs due to the lack of a widely accepted standards or methods for public sector ROI analysis. Without such standards, governments have difficulty obtaining and using the kind of "bottom line" information that could reveal the value of IT investments across all kinds of programs and help guide new investments. These difficulties have not, however prevented many government agencies from including some public return considerations into IT investment and development decisions.

This study describes how public ROI considerations entered into the development and implementation of five cases of government IT investments, located in Canada, Europe (Austria), Israel, and the US. The kinds of IT investment vary considerably across the cases, including a state level digital archive for government records in Washington, a government-wide enterprise resource planning application in Pennsylvania, a shared application and infrastructure system for public services in New Brunswick, a tax administration and collection system in Austria, and financial management and purchasing system in Israel.

The case descriptions are based on interviews conducted during 2006 with participants in these cases and examination of related documentary material. In addition, the design for the case studies and initial exploration of public return assessment included an agenda-setting workshop that included twenty government officials and researchers from Canada, Europe, and the US. The results of the workshop have guided the framing of the research and identification of case study sites. At this writing three of the

five case studies have been completed and two are underway. The results will be available for the full poster presentation.

A core issue in these cases is how the government agencies conceive of and describe public ROI assessment. In the absence of standards the number of definitions and components in a public ROI assessment can be very large, since it can include both costs and returns in social, political, and economic terms that are broader and softer than the hard financial measures used in business. Of course, financial measures are used in government, but they seldom represent the full range of returns generated from public investments in IT. The cases explore how the decision makers developed and adapted various elements of public returns in their planning and justification for the IT investments. The case descriptions also include attention to how the agencies collected evidence of investment results and returns, as well as how those results were communicated with stakeholders.

2. FRAMEWORK FOR PUBLIC ROI ANALYSIS

The case studies and workshop results will be used to complete an analytical framework for conducting public ROI assessments in government IT projects. Preliminary work on that framework will be included in the poster presentation. These will include the schematic in Figure 2 (below). It identifies the relationship of traditional ROI value linkages to the broader value proposition involved in the more comprehensive public approach.

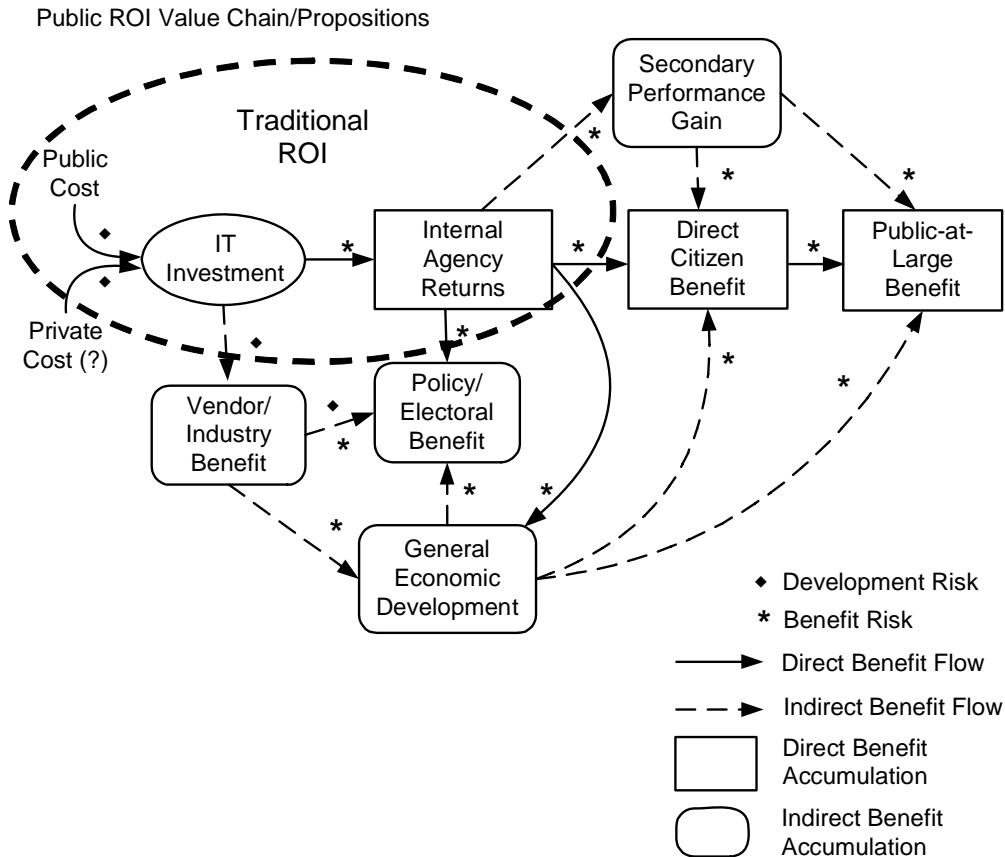


Figure 1. Public ROI Value Chain/Propositions

- *Development risk* – risk to the prospects of generating public value that is a consequence of possible failure of the technology development effort itself.
- *Benefit risk* - risk to the prospects of generating public value resulting from other factors preventing or reducing public benefits even if the technology development is a success.
- *Direct benefits* are those that accrue to the public through a mechanism that links the beneficiary to the government activity/program without a mediating person or agency (e.g., the local government plows snow on my street.)
- *Indirect benefits* are those that accrue to the public through a mechanism in which the government action affects an

intermediary person or agency, which in turn benefits the public (e.g., The State of New York contracts with United Way to operate homeless shelters).

- *Secondary Performance Gains* – refers to performance enhancements resulting from the IT investment that occur in a separate government agency or unit, not directly involved in the initiative—a spillover benefit that enables another agency to improve its own performance.

The full framework will include guidance on choosing assessment methods, analysis, and presentation of results. Separate reports of each case and a comparative analysis will be included.

Eco-Informatics and Natural Resource Management

Judith Bayard Cushing
The Evergreen State College
Olympia WA 98505
USA
01-360-867-6652

judyc@evergreen.edu

Tyrone Wilson
National Biological Information Infrastructure
U.S. Geological Survey
Reston, VA, USA
01-703-648-4075

tyrone_wilson@usgs.gov

Alan Borning (University of Washington), Lois Delcambre (Portland State University), Geoff Bowker (Santa Clara University), Mike Frame (USGS/NBII), John Schnase (NASA), William Sonntag (EPA), Janos Fulop (Hungarian Academy of Sciences), Carol Hert (University of Washington-Tacoma), Eduard Hovy (University of Southern California), Julia Jones (Oregon State University), Eric Landis (Natural Resources Information Management), Charles Schweik (University of Massachusetts-Amherst), and Lawrence Brandt, Valerie Gregg, and Sylvia Spengler (National Science Foundation).

ABSTRACT

This project highlight reports on the 2004 workshop [1], as well as follow-up activities in 2005 and 2006, regarding how informatics tools can help manage natural resources and decide policy. The workshop was sponsored jointly by sponsored by the NSF, NBII, NASA, and EPA, and attended by practitioners from government and non-government agencies, and university researchers from the computer, social, and ecological sciences.

The workshop presented the significant information technology (IT) problems that resource managers face when integrating ecological or environmental information to make decisions. These IT problems fall into five categories: data presentation, data gaps, tools, indicators, and policy making and implementation. To alleviate such problems, we recommend informatics research in four IT areas, as defined in this abstract and our final report: modeling and simulation, data quality, information integration and ontologies, and social and human aspects. Additionally, we recommend that funding agencies provide infrastructure and some changes in funding habits to assure cycles of innovation in the domain were addressed.

Follow-on activities to the workshop subsequent to dg.o 2005 included: an invited talk presenting workshop results at DILS 2005, publication of the workshop final report by the NBII [1], and a poster at the *NBII All Hands Meeting* (Oct. 2005). We also expect a special issue of the JIIS to appear in 2006 that addresses some of these questions. As we go to press, no solicitation by funding agencies has as yet been published, but various NASA and NBII, and NSF cyber-infrastructure and DG research efforts now underway address the above issues..

Categories and Subject Descriptors

H.2.8 [Database Applications]: Scientific Applications.

Keywords

Information integration, modeling and simulation, data quality, human-computer interaction, web services, eco-informatics.

1. INTRODUCTION

The 1998 PCAST report characterized bio-informatics as a biology and CS/IT cross-discipline, recognized the biodiversity-ecosystem nexus as an information enterprise, and envisioned analytical and synthetic capabilities among other foci in the next

generation of NBII-2 information services [2]. Informatics tools for solving environmental challenges, e.g., global climate change, emerging diseases, decreasing biodiversity, and waning resources, are researched and developed under the *eco-informatics* rubric, but IT research supporting natural resource management remains largely absent from the academy. Participants in the 2004 Eco-Informatics and Policy Workshop sought to address this gap.

2. THE PROBLEM SPACE

The problem space and required IT for eco-informatics policy-making is broad because stakeholders include communities of interest as well as of place. Further, political aspects of the data necessitate equitable information access and mass customization, and distinguishing among measurements, indicators, and interpretations. Finally, metadata and validation are essential. More specific problems are categorized into five areas:

1) Data presentation problems arise from complex interactions between user needs and data – metadata, raw data, accuracy specifications, methods, documentation, and policy. Needed research includes determining what information works best on which medium, cross-referencing and supporting data across presentations, representing time and change, using new media, e.g., 3D or VR, and user task definitions. This area distills into: presentation as mediator among users, and data-metadata issues.

2) Geographic data gaps stem from a lack of needed data sets or access to them, disjunct data sets requiring manipulation because of temporal or spatial gaps, an emphasis on adaptive management that out-paces data reliability, and the lack of IT professionals upon whom resource managers can call for expertise. Major issues include how to generalize fine-scale data containing gaps, and decision makers' and policy makers' sensitivity to uncertainty.

3) Tool problems involve a) lack of a tool “clearing-house”, b) problems of data collection and new or different data types, c) lack of user frameworks and development standards, d) lack of tools to support metadata issues, and e) social science issues of usage, sharing, and adoption.

4) Indicator problems exist because indicator definition, relevance, and value are neither well-defined nor well-communicated. Constituents may be uneasy with environmental measures, and data gaps adversely affect the reliability and trust that stakeholders place in indicators. The inherent complexity of ecosystems further complicates this issue.

5) Policy and implementation problems include, but are not limited to, problems organizations encounter in the: a) financing, production, and maintenance of tools and information; b) use and possible abuse of tools or information, and determining their effectiveness; c) cross-organizational sharing or not sharing, e.g., privacy, confidentiality policies, of IT tools or information; d) communication, or lack thereof, of environmental management decisions grounded in eco-informatics-based analysis.

3. RESEARCH ISSUES

To articulate research issues, interdisciplinary approaches adopted by several current successful research projects were examined. Strategies for sustaining future research and case histories that exemplified the need for the research were identified as critical. Four research areas were identified. Research areas defined below cut across the problems identified in Section 2 above.

1) Modeling and simulation research issues included: coupling diverse models, exploring model design values for diverse stakeholders, incorporating visualizations with model results, representing error and uncertainty, handling large data sets, and open source modeling infrastructure. An open-source, flexible, reusable modeling infrastructure, along with the social practices that sustain it, were seen as critical since it would allow researchers and decision makers to experiment freely with new models or change existing ones.

2) The problem of data quality is how to determine and communicate uncertainty to decision-makers for studies integrating multiple data sources. Methods are needed to mitigate error when creating and combining data sets, and to associate error with alternative decisions. NSF could publish metadata standards for all grants, rather than just certain programs, and make metadata obligatory for all data sets. Whether uncertainty associated with data synthesis influences resource management, whether decision-maker perceptions of the value of findings from synthesized data, and how synthetic studies compare in courts of law to other “expert testimony” remain open questions.

3) Information integration involves mechanisms for reliable, transparent and authoritative data combination. Associated research includes: defining the dimensions of integration; quantifying semantic distance; integrating multiple ontologies; promoting document modeling; evaluating utility of qualitative and quantitative data; tools to support data integration; and evaluating knowledge from non-traditional sources. **Ontologies** are useful in providing semantics over databases, making cross-disciplinary connections, and producing thesauri. Tools to build, verify and deliver ontologies require considerable research. Other research includes understanding gaps and inconsistencies in ontologies, trust and verification of ontology content, and ways to handle change in ontologies that go beyond versioning.

4) Social and human aspects of eco-informatics and policy include: broad collaboration in IT tool development and information sharing, measuring success and determining appropriate institutional designs and incentives or disincentives; human-computer interaction; management practices, graduate education and data management training. Advancing the eco-informatics agenda hinges both on new technologies and new understandings of relationships among individuals, organizations, communities, disciplines, information resources, and tools.

4. RECOMMENDATIONS

Workshop participants recommended funded research to address issues in Section 3, and emphasized that attention be paid to assuring innovation cycles from research to prototype, to development and commercialization, to deployment and evaluation, and back to research. This is critical because non-overlapping missions and reward systems of different agencies make it easy to lose momentum. Research funding longer than three years is needed to elicit requirements, integrate them into research, and engage an “agile” software cycle with stakeholders.

This domain is similar to digital government IT research in that finding the right domain problem, distilling a range of research fruitful to all stakeholders, finding the right agency collaborators and managing expectations, are all critical. Funding agencies should work together and with principal investigators, information managers and decision makers to sustain and encourage innovation, research and development, and education and training. To be addressed in these collaborations include: 1) Matching researchers and agency collaborators, 2) Bringing research results and prototypes to resource managers in field offices, 3) Evaluating prototypes and products to understand strengths and weaknesses and inform new research, 4) helping researchers understand how to combine quantitative with qualitative information, and decision-making processes.

Further, considerable attention must be paid to constant reprioritization of the research agenda, since the number, breadth and complexity of problems and potential solutions suggested herein dictate decades of research – while species and ecosystems disappear at an increasing rate. So, assuring the development of tools that through extensibility promise application to a wide range of problems is critical. Furthermore, it will be important to keep a range of research projects in the pipeline – from highly theoretical and generalized, to working prototypes developed by researchers and resource managers, to deployment experiments.

5. ACKNOWLEDGMENTS

The authors thank presenters¹ and participants in the Eco-Informatics and Policy Workshop, as well as Aaron Ellison of Harvard University and Elaine Hoagland of The National Council for Science and the Environment for helpful comments. Finally, we acknowledge funding from the National Science Foundation (NSF IIS 0505790).

6. REFERENCES

- [1] The workshop website includes all presentations, the final report, and participants <http://www.evergreen.edu/bdei>.
- [2]<http://clinton3.nara.gov/WH/EOP/OSTP/Environment/html/teamingcover.htm>

¹ Additional presenters include: C. Marie Denn (USPS), Richard Guldin (USFS), Stephen Jensen (EU), Paul Klarin (Oregon Coastal Management), Ron Li (Ohio State University), Molly O’Neil (Harry Reid Center for Environmental Studies), Phil Rossignol (Oregon State University), Mark Simonson (Puget Sound Regional Council), Larry Sugerbaker (NatureServe), Nancy Tosta (Ross&Associates), Dawn Wright (Oregon State University).

Overview: Building a Sustainable International Digital Government Research Community

Sharon S. Dawes
Center for Technology in Government
University at Albany/SUNY

sdawes@ctg.albany.edu

1. ABSTRACT

This overview describes a four-year effort to create a framework for a sustainable global community of practice among digital government researchers and research sponsors. The project will support an international reconnaissance study describing the current status of digital government research, an annual research institute, a framework for several international working groups, and travel support for US investigators and doctoral students to participate actively in international conferences and workshops. Project results will be disseminated widely using a variety of publication and communication channels. The project will run from late 2005-2009.

Keywords

Community of practice, international collaboration.

2. An Emerging Global Research Domain

Over the past decade, growing evidence demonstrates the emergence of a global field of inquiry at the intersection of government, society, and information and communication technologies. This domain is often characterized by “e-government,” “e-governance,” “information society,” and other related terms. We use the term “digital government” to encompass this collection of research ideas. In the United States (US), the National Science Foundation’s (NSF) Digital Government (DG) Research Program has provided leadership and support for this relatively new domain of research. In Europe, the European Commission, as part of its Information Society Technologies (IST) program, sponsors an ambitious e-government research program. At the same time, the research councils of individual European states support comparable research programs within their borders. Similar efforts are established or emerging in Canada, Australia, India, the Pacific Rim, Latin America, and Africa. International organizations such as the United Nations and World Bank support e-government development and are also becoming interested in associated research.

Because of the relative newness of the DG field, there is insufficient interaction among researchers in different countries compared to what one finds in more established scientific disciplines. As this is a relatively new domain of inquiry, it involves multiple disciplines (a challenge within a single country, let alone internationally) and there are very few support mechanisms and forums to engage DG researchers with their peers working in this domain around the globe. Furthermore, once a potential collaboration starts that could lead to joint research efforts, it is logically and financially difficult to

Valerie Gregg
Digital Government Research Center
Information Sciences Institute, USC

vgregg@isi.edu

sustain it to the point of joint research proposals and reliably funded projects. Consequently, comparative and transnational issues in DG, which are of growing importance in an increasingly networked world, are not receiving the attention they deserve.

3. Objective: An International Digital Government Research Community

Trends in digital government research and the limited international experiences gained so far suggest at least three ways to internationalize investigations and bring the benefits (and the challenges) of multi-cultural perspectives to this important worldwide field of research:

- Create opportunities for scholars interested in particular domains of study to encounter the work of international colleagues and to engage in discussions that can lead to shared research agendas and joint projects, as well as the more traditional exchange of individual methods and findings.
- Encourage the investigation of international problems that governments routinely must address, such as drug interdiction, immigration, global trade regulation, or border control.
- Support comparative studies that seek universal theories and transferable practices by studying selected phenomena in a variety of cultural settings using consistent designs and methods, with explicit points of comparison and evaluation.

This project focuses primarily on the first item as the means to achieve the second and third. Our strategy has several mutually reinforcing streams of work as follows:

3.1 International Digital Government Research Review (Reconnaissance Study)

A reconnaissance study will identify and summarize the state of international DG research. The results can be used as a baseline benchmark for assessing its subsequent growth and development. The study can also inform the development of a global research network and associated comparative and transnational projects in the digital government domain. It will rely on interviews, literature reviews, and documentary analysis and will address questions such as the following:

- What international problems are or have been the subject of digital government research efforts? What has been learned?
- What topics have been investigated using comparative methods across national boundaries? What has been learned?

- What problems and topics are or have been emphasized by different research sponsors?
- What are the patterns of investigation (problems, topics, methods, funding sources and mechanisms) in different parts of the world?
- What are the important international organizations in this research area? Who are the principals?
- What are the most important research institutions, conferences, journals, blogs, or other online sources of research information that span countries?

3.2 Developmental Working Groups

Following review with the advisory group and public dissemination of the preliminary reconnaissance study, the investigators will organize a competition to select and provide support for topical working groups. Three to five groups will be selected by peer review. We expect each group will involve about 12- 20 people, including both established researchers and doctoral students. For each group selected, travel support will be provided from NSF funds for US participation in five focused working group meetings over the course of three years, including travel support for at least two US doctoral students in each group. (Travel by participants from other countries needs to be supported by other sponsors). We believe the working groups are an excellent way to introduce doctoral students to international research issues and programs. We expect the relationships formed in these venues will be long-lasting professional connections that serve to further strengthen the network of international digital government researchers.

The expected result from the first working group meeting in each topical area will be a formal research agenda including both comparative questions (i.e., problems that occur in multiple countries) and transnational questions (i.e., problems that are international in scope). Over the course of two years, the members will be in routine communication, co-author journal articles, participate in international conferences,. Following the final meeting, each group will produce a white paper authored jointly by the US and international participants that discusses the research challenges, recommended strategies for undertaking this research, and the accomplishments within its sub-domain.

3.3 Annual Research Institute

Each year, an international institute on digital government research will provide an intensive residential program for comparing research themes, methods, and results, as well as for building a deeper mutual understanding of the multi-disciplinary nature of DG research. In the first year, a small number of experienced DG researchers from different countries and

disciplines will come together in a week-long faculty-only residential institute. Their goal will be to familiarize one another with their fields and research approaches and to jointly design an annual institute program for doctoral students that would begin in the second year. We expect the program design will address such topics as (1) explicit comparisons of the philosophies, questions, and methods among the disciplines that make up digital government research, (2) a review of pressing comparative and transnational research questions and ways to study them, (3) how to design an international investigation (4) how to manage an international project, (5) how to apply multi-method and multi-disciplinary approaches, etc. This grant would provide a modest honorarium plus travel and residential expenses for five U.S. faculty per year. In the second and third years, it would also provide scholarships for up to ten U.S. doctoral students. International faculty expenses and similar student scholarships will need to be covered by other sponsors. During the third year, we would develop a business plan for making the institute self-sustaining. By the forth year, we expect the program could become self-funded through tuition and fees.

3.4 US Participation in Conferences

In addition to costs of US participation in the working groups and annual institute, this NSF grant will support travel costs for 4-8 US participants annually to take active part in international research conferences that place a substantial focus on digital government research. To receive support, US participants must have a leadership role in international research and/or have accepted papers or panels in which they are co-authors or co-presenters with international colleagues. A larger number of trips would be supported in the early years to jump-start relationships that we expect will lead to other funding for travel in later years.

4. Expected Outcomes & Dissemination Plans

By the end of this project we expect the connections and collaborations developed among investigators and research sponsors to result in new knowledge discoveries and new lines of intellectual inquiry. Further, we expect to have made significant contributions towards establishing a baseline benchmark and periodic assessments of international DG research efforts and results; created new mechanisms for training top doctoral students in the interdisciplinary and international field of DG; supported several international sub-communities (the working groups) in DG research as well as a replicable model for additional groups; created awareness and initial commitment to international research collaborations among research sponsors and, enabled many face-to-face informal relationships between US and international researchers that can be the foundation for long-lasting collaborations.

Technology Adoption and Institutional Change in the United States Senate: An Analysis of Web Site Content

Kevin Esterling

UC Riverside

900 University Ave.

Riverside, CA 92521

951-827-3833

kevin.estrling@ucr.edu

David Lazer

Harvard University

79 JFK Street

Cambridge, MA 02138

617-495-1100

David_Lazer@harvard.edu

Michael Neblo

Ohio State University

154 N Oval Mall

Columbus, OH 43210

614-292-6446

neblo.1@osu.edu

1. ABSTRACT

This poster examines data gathered from the official websites of the US Senate to examine whether technology enhances online representative functions.

Keywords

US Senate, Websites, Technology Deliberative Democracy, Structural Equation Modeling

1. INTRODUCTION

For many observers, the rapid emergence of the public sector's Web presence has raised expectations for institutional change and improvements in democratic governance [7]. As government organizations adopt new information technologies, the logic goes, public organizations will themselves adapt to the technical expectations of all participants in the information-based economy, an economy that places a premium on rapid and free information flow, innovation, and analytical thinking [2, 3].

One of the key set of public organizations in a democracy is the legislature. In this poster, we examine how the U.S. Senate has adapted to the opportunities that the Internet offers. The Web presence of the U.S. Senate has undergone a dramatic change over the past decade, from a handful of Senators establishing individual gopher sites in the early 1990s to the current Senate Web homepage (www.senate.gov) that contains official Web pages for every Senator [1, 5]. This development has the potential for increasing informed involvement among citizens, interest groups, and social movements in the Senate's legislative process. This paper examines whether and how the Senate has actually utilized their Web presence consistent with this potential.

The study of U.S. Senate Websites is interesting and important if only because of the central role the Senate plays in constitutional democracy. In addition, legislative institutions provide a critical test case of the expectations for the transformative role of communication technology in the public sector, given the normative importance of information, innovation and analytical thinking in legislative settings. If Web-based communication technologies indeed possess transformative capacities, this information technology itself can enhance the possibilities for deliberative forms of democratic representation in the Senate. To

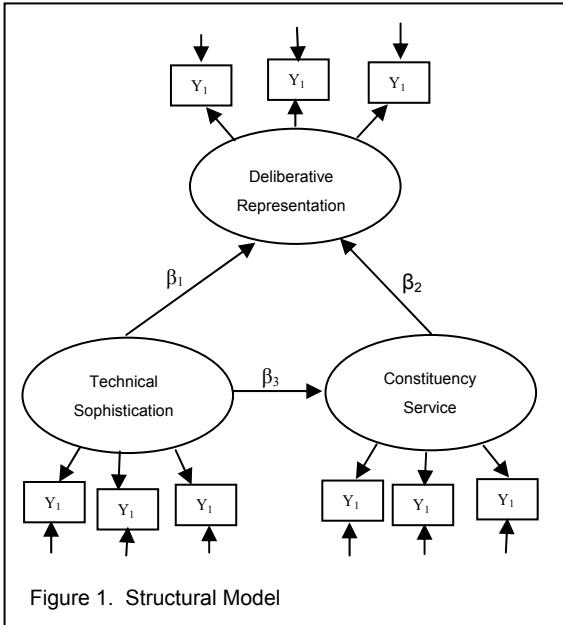
the extent this transformation occurs, legislative outcomes better approximate normative standards of democratic theory.

As Jane Fountain [3] notes, however, adopted technologies and political institutions are mutually constitutive. While it is certainly conceivable that Web-based technologies can improve information flow and citizen participation in the legislative process, it is equally true that the incentives and norms embedded in legislative institutions will shape how Senate offices use technology. Technology is not necessarily used in an optimal fashion. Instead, institutions may provide incentives to adapt new technologies to existing standard operating procedures and organizational norms, and create disincentives to take risks with innovations [2, 3]. In the case of Congress, descriptive studies have well-documented the tendency of Congress to coerce new technologies into old practices, such as printing out hard copies of constituent email to be processed with the postal mail, and using the official Web page to reproduce printed pamphlets [5]. As Fountain [3] argues, "the major challenge for government is not the development of web-based [government-to-constituent] transactions but reorganizing and restructuring the institutional arrangements in which those transactions are embedded" (page 6).

2. STATISTICAL TEST

In this poster, we statistically test these competing "optimistic" and "pessimistic" expectations regarding the potentially transformative role of Web communication technology on the representative functions of the U.S. Senate.

We use data collected by the Congressional Management Foundation (CMF). In 2002, CMF coded a wide range of attributes of the Websites of all Senators, relating to the relative technical sophistication of the sites, features that enhance constituency services, and the relative informational content of websites regarding local and national issues before the Congress. Our research questions focus on whether the technical sophistication of websites enhances, detracts, or is unrelated to the representative functions of websites. The optimistic view of websites asserts that technology plays a positive transformative role, and so expects that increasing technical sophistication of websites should tend to be associated with improved representative functions. The pessimistic view holds that Senators tend to coerce technology to fit existing standard operating procedures.



To test these expectations, we use structural equation modeling (SEM). In the SEM statistical framework, coded attributes of websites serve as indicators of unobserved or unmeasured latent factors that are of substantive interest. Simultaneously, latent factors can be regressed on other latent factors in a structural model. In the SEM for this paper, we specify three latent factors: technological sophistication, constituency services, and deliberative representation. See Figure 1.

The indicator variables are taken from the CMF coding dataset (for a description of these codes, see Johnson 2004). We measure technological sophistication with four variables that indicate the relative comfort and sophistication of the member's staff with web-based technology: the ease with which the coder could navigate the website, the timeliness of the information content of the website, the readability of the content, and the scannability of the website. All of these indicators are measured 1 to 5e.

Owen et al. [5] and Adler et al. [1] note that members of Congress make extensive use of their websites to provide services for constituents. To measure constituency service we include indicators of the extent to which the site provides links to agencies relevant to constituents' interests, instructions for how a constituent may initiate case work with the member's office, and the extent to which the site provides case work FAQs (frequently asked questions). All of these indicators are measured 0 to 5.

We also wish to measure a latent variable for the extent to which the website conforms to normative standards for deliberative representation. Deliberation requires that the member allow her opinions to be subjected to public scrutiny. As Wilhelm [7] notes, in deliberative democracy, "interlocutors in a political debate need to provide reasons to support their arguments, reasons that can be validated intersubjectively" (page 43). To measure the extent to which websites reach the normative ideal of providing issue rationales, we include measures of the presence of issue content of importance to the member, of importance to the nation, and of importance to the district (all 0 to 5).

We assume the latent variables are distributed multivariate normally each with mean zero, and use a structural equation model to estimate their variances and covariances. We specify an ordered logit link function for the measurement models regressing the indicators on the latent variables, estimating a separate set of thresholds for each indicator (for a total of 46 threshold parameters and 7 free factor loadings). The β s capture the magnitude of the relationships among the latent variables.¹

3. RESULTS AND CONCLUSIONS

The SEM estimates relationships among the latent variables in the model. Consistent with the "pessimistic" institutionalist view of technology adoption, we find that both constituency service and deliberative representation are *negatively* correlated with technological sophistication. Senators who have websites that demonstrate a greater level of comfort with web technology (those sites that are easier to navigate and use) tend to have worse constituency services and lower levels of deliberative representation. In addition, the results suggest there is no relationship between the constituency service orientation of Senate websites and any emphasis on deliberative representation.

Taken together, these preliminary results suggest that Senators, like other public authorities and actors [3] tend to coerce technologies to fit their institutional needs. Members who tend to use their websites to perform traditional representative functions do not appear to be influenced by the novel fact that this function is being performed online. Conversely, members who demonstrate comfort in using technology appear to focus on the technological means, that is, having technically sophisticated websites, rather than on the ends to which the technology should be used, that is, to improve representative functions and democratic governance. These findings echo the arguments of the e-government institutionalists that often the greatest difficulty in improving e-government is in changing the underlying culture.

4. ACKNOWLEDGEMENTS

This project is generously funded by a grant from the National Science Foundation, Program on Digital Government (NSF 001684-002).

4. REFERENCES

- [1] Adler, E. Scott, et al., "The Home Style Homepage," *LSQ* 23(4): 585, 1998.
- [2] Dawes, Sharon et al. "Some Assembly Required," SUNY Center for Technology in Government, 1999.
- [3] Fountain, Jane E. *Building the Virtual State*, Brookings, 2001.
- [4] Johnson, Dennis W. *Congress Online*, Routledge, 2004.
- [5] Owen, Diana et al., "Congress and the Internet," *Press/Politics* 4(2): 10-29, 1999.
- [6] Roscoe, Timothy, "The Construction of the World Wide Web Audience," *Media, Culture and Society* 21:673-84, 1999.
- [7] Wilhelm, Anthony G. *Democracy in the Digital Age*, Routledge, 2000

¹ In this preliminary analysis, we have not estimated the structural regression coefficients among the latent variables (the betas in Figure 1).

eGovernment for Business across the Atlantic: from Cases to Models

Patrizia Fariselli
NOMISMA

NET-Net Economy & Technologies
Strada Maggiore 44
+39-051-301026
farisellip@nomisma.it

Julia Culver-Hopper
NOMISMA

NET-Net Economy & Technologies
Strada Maggiore 44
+39-051-6483181
culverj@nomisma.it

Olana Bojic
NOMISMA

NET-Net Economy & Technologies
Strada Maggiore 44
+39-051-6483133
bojico@nomisma.it

Keywords

eGovernment Services - Technologies - Strategy - Integration;
eGovernance; Trans-Atlantic Networking

This paper presents the methodology and conclusions of a study - subcontracted by CIMIC-Rutgers University within a project funded by the NSF dg.o program - aimed at comparing a CIMIC-New Jersey (USA) project for integrating online Public Services (ePSs) for businesses with similar cases based in the European Union (EU). The study attempts to gain a general understanding of the key issues associated with intra & inter-Public Administration (PA) agency governance, case sustainability, and networking that are as critical in the USA as in the EU by analyzing a multitude of individual eGovernment cases at local, national or European levels. Another interesting direction to be explored concerns the networking across the Atlantic of dedicated portals for businesses. Though the cyberspace is essentially global, normally the one-stop shops of ePSs target business communities which are 'local', no matter they are also national, or even international. As transnational linkages should be further developed within the EU, leveraged by pan-European eGovernment services and initiatives aiming not only to diffusion, but also to strengthen European integration, the same expectation should be fostered in a transatlantic perspective.

The project undertaken by CIMIC in New Jersey [1] prepares the ground for delivering improved, faster and integrated administrative services for new businesses to be established within the State. The eGovernment project is instrumental to the State policy aimed at accelerating private sector investment, improving the attractiveness and competitiveness of the State as a location for business. The project involves both public sector and academic partners and benefits from a NSF matching grant.

The objective of the study granted to Nomisma-NET within the NSF-CIMIC project was to determine the state-of-the-art of provision of integrated ePSs for businesses in the European Union - comparing them with the CIMIC-New Jersey case – in order to identify differences and similarities in the US and EU approaches and models of eGovernment. Since it is extremely difficult to find perfectly comparable cases, even within the same country, a formal comparison of cases outside their context is neither expected to be possible nor desirable. Our intent, therefore, was to

produce an analytical framework starting with the CIMIC-New Jersey case in order to use it as a methodological reference when navigating in the open sea of EU eGovernment for business cases, in order to identify any emerging model from the multitude of EU cases. Currently, in the EU the analyses consist mostly of inward-looking case studies, or horizontal surveys of cases categorized by specific technical features, or as 'good practices' based upon parameters of the ePS supply in the EU Member States [2]. Since the current policy in the EU [3-4] gives priority to the diffusion of eGovernment take-up across the various administrative levels, the measurement of the supply prevails over the measurement of case sustainability and impact on demand. For these reasons, the identification of any EU eGovernment model becomes an *ex post* exercise, requiring an external reference. The CIMIC-New Jersey case plays exactly this role. The research has been articulated in the following steps:

1. Identification of units of analysis

In the EU, as in the USA, the geo-administrative units creating eGovernment aimed at promoting business development are highly varied and range from the municipal level through the regional and national levels. Therefore, it results as extremely difficult if not useless to try to achieve perfect symmetry with regard to the geo-administrative units of comparison between the New Jersey and EU cases. We thus have examined cases at all levels of administrative authority, focusing on the supply of integrated ePS for businesses rather than on the administrative level of the supplier.

2. Determination of case criteria

SERVICES
TECHNOLOGY
STRATEGY
<ul style="list-style-type: none">• Case Implementation & Coordination• Network Governance
INTEGRATION
<ul style="list-style-type: none">• Technical Integration• Policy Integration

In our approach, integration of services reflects both the degree of efficiency of the eGovernment supply and the degree of internal cohesion and sustainability of the individual case and of the larger PA network, which only can be ensured by a governance model based on participation and partnership among PA officers, technology providers, and organization experts. The effective organization and delivery of ePSs requires thorough attention to information management in the back-office, so that efficient and sustainable ePSs integration occurs only if new services are supplied based on the innovative recomposition of public information.

3. eGovernment for Businesses in the EU

Realization of eGovernment is one of the objectives of the Lisbon Strategy (launched in 2000). Under eEurope 2005 program [5], EU governments committed themselves to putting 20 services online (12 aimed at citizens, 8 at businesses) by the end of 2002. The achievement of this Lisbon target is measured by benchmarking the level of online availability and sophistication of the eGovernment services in Europe according to a four-stage maturity curve [6] based on the degree of electronic interactivity between PA, citizens and businesses. The service-centred EU policy is strongly oriented to the supply side and to the efficiency gains to be achieved by the PAs through eGovernment applications, infrastructures and reorganisation models, rather than to demand and to intensity of information of services.

The primary building blocks of EU policy for eGovernment aimed at businesses are associated with the following concepts:

- *one-stop shop for business*

One of the most important policy developments since 1990s in the area PA2B was the adoption in various European countries of the Single Access Point for Business ('one-stop-shop') organizing several different services in a single user interface office or portal. While originally represented by a physical structure this concept has then been translated into eGovernment applications.

- *back-office reorganisation*

The Back-office Reorganization study [7] was a seminal work in examining the reorganization induced by the adoption of eGovernment, impacting on the PA workflows and on the integration between front & back offices involved in an administrative process. Integration is defined in a supply chain perspective, as integration of horizontal services, sub-processes, data.

4. Overview of selected case studies

Two categories of portal came under analysis:

- a) Single purpose portals
- b) Integrated services portals
 - Portals for Businesses
 - Integrated Portals for Citizens and Businesses

The study also provides a general overview of the uptake of eGovernment services for business at the national level and the level of sophistication of those of the 8 ePSs which are supplied also by the CIMIC-New Jersey project.

5. EU Model vs Reference Model

The EU eGovernment policy shows a supply/service/process-oriented approach to Integration, implemented through portals incorporating the 'one-stop shop' concept for businesses and citizens. Integration occurs along two parallel channels: one focusing on the front-office (indicator: the degree of interactivity between PA and users along the maturity curve), the other involving the back-office (indicator: the degree of interactivity and interoperability within the PA agencies). Cases self-description and literature put emphasis on Services and Technology, while Integration is mostly associated with the technical issues of interoperability within and between PA agencies. Technology is given the responsibility to transform fragmented PA organizational setting into a seamless supply chain. Almost no insights are provided on Strategy, except descriptions of the technical steps for implementing the platforms. What emerges is a supply-driven model, aimed at reforming the

European PAs by re-engineering the administrative process through the delivery of existing PSs via electronic channels, and by the aggregating ePSs in gateways interfacing the PA and users.

When comparing the EU model with the reference model drawn up from the CIMIC-New Jersey project, differences and similarities emerge. Basically, the two models share similarities as far as Services and Technology are concerned; thus, comparative analysis is technically possible only for these two criteria on a case-by-case basis. The contrasts are due to different:

- *purposes*: stimulating business competitiveness vs. PA harmonization and efficiency; one U.S. State vs. many EU States
- *approaches*: supply vs. demand-driven eGovernment policy; general framework vs. operational cases; ePSs + 'intelligence' to target users vs. e(basic)PSs to all
- *emphasis*: open vs. intermediate access public information; information-intensive ePSs vs. interactive ePSs.

The major divergences between the reference model and the EU model concern the Coordination Strategy towards Integration, as illustrated in the table below.

REFERENCE MODEL	EU MODEL
PA + Tech + Biz Coordination	Tech → PA → Biz Transfer
PA2PA before ePSs	PA2PA after ePSs
Linear process to Integration	Spiral process to Integration

The EU model extensively applies to eGovernment typical concepts of the business management disciplines, such as productivity, efficiency, value chain management, time-to-market, customer relation and satisfaction. Yet, the model, opens up four major trade-off about Costs, Governance, Demand, Information. What emerges from recent researches [8] and surveys [9] is that the business demand of public information online exceeds the demand of electronic transactions for getting ePSs. Investment in intensive-information based ePSs represents the most fruitful direction to eGovernment, exploiting the deep potential of digital-network technologies for creating new services, rather than only transforming online the present ones or pooling them onto a single access point

REFERENCES

- [1] Adam N., Atluri, V., Chun, S., Fariselli P., Culver-Hopper J., Bojic O., Stewart R., Fruscione J., Mannocchio N, Technology Transfer of Inter-Agency Government Services and their Transnational Feasibility Studies, Proceedings of the NSF dg.o 2005 Conference, Atlanta, Georgia, May 15-18, 2005.
- [2] European Commission, DG Information Society and Media, *A Good Practice Framework*, www.egov-goodpractice.org
- [3] European Commission, *The Role of eGovernment for Europe's Future*, COM(2003) 567 final, 26 September 2003
- [4] European Commission, *eGovernment Beyond 2005 - Modern and Innovative Public Administrations in the 2010 horizon*, 3rd eEurope eGovernment subgroup meeting, Amsterdam, NL, 27-28 September 2004
- [5] European Commission, *eEurope - an Information Society for All*, COM(2000)130, 8 March 2000

- [6] Cap Gemini Ernst & Young, *Web-based Survey on Electronic Public Services, Results of the Third measurement, October 2002*, 2003
- [7] DTI-Danish Technological Institute, IFIB-Institut fuer Informationsmanagement Bremen GmbH, *Reorganisation of Government Back Offices for Better Electronic Public Services - European Good Practices (Back-office Reorganisation)*, Final Report to the European Commission, 2004
- [8] Fariselli P. (ed.) *Tecnologie dell'informazione e imprese* [Information Technologies and Enterprises. Demand and Supply of Public Information Online in Italy], Nomisma Studi e Ricerche, 2005
- [9] Eurostat, *e-Government: Internet based interaction with the European businesses and citizens, Statistics in Focus*, 9/2005

SGER: Project Summary - CAPWIN

Mark Gaynor

1st author's affiliation
595 Commonwealth Ave
Boston, MA, 02215
01-617-353-4159
mgaynor@bu.edu

ABSTRACT

Project highlight for CapWin NSF SGER grant.

Categories and Subject Descriptors

D.3.3 [Programming Languages]: Standardization, Experimentations.

Keywords

EMS services, Standards.

Our project is focused towards building an emergency medical response application based on emerging standards and technologies that will interoperate with governmental departments such as DOJ, consortiums of governmental agencies within a regional boundary such as CapWIN and the military. Our application called iRevive is a sensor-supported, pre-hospital patient care system that allows flight nurses and paramedics to capture and transmit electronic patient care data from the field to one or more central locations in near real-time. It employs several different wireless devices and adheres to a number of emerging standards for the storage and transfer of electronic patient data. The intelligence of the system is in its use of knowledge-based rules to dynamically generate a wide variety of illness-specific protocols and forms for clinical documentation. By capitalizing on several new and emerging technologies, iRevive offers a robust, flexible, and extensible information technology infrastructure for complete and accurate patient data collection in the pre-hospital phase of patient care.

We have had challenges and successes including a reorganization of CapWIN, one of our major partners, and the addition of the Dept of Justice, and Boston MedFlight as partners. CapWin has undergone a change of management, and a re-focusing of its primary chat room application. We are still working with CapWIN for a long term deployment of our EMS application. We have started discussions with DOJ on the applications of sensor networks. Our major success is with our partner Boston MedFlight, we are developing their next generation patient documentation application built on emerging standards that will promote interoperability between governmental departments.

There are four phases of a typical mission: pickup, transport, drop-off, documentation completion and data transfer. Data capture begins when the transport vehicle (helicopter, jet or ground ambulance) arrives at the patient pick up point, which is either a medical facility or injury scene. iRevive is used by nurses and paramedics to enter data, starting with the patient's

current condition. As transport commences, emergency medical specialists treat the patient and time permitting, initiate the documentation process. The data capture phase ends when patient care is transferred to the care of physicians and nurses at the receiving hospital. At this point patient data can be transferred into in-hospital medical IT systems. The final phase requires completing all necessary documentation of the mission and transferring this data to a database for billing, storage and future retrieval.

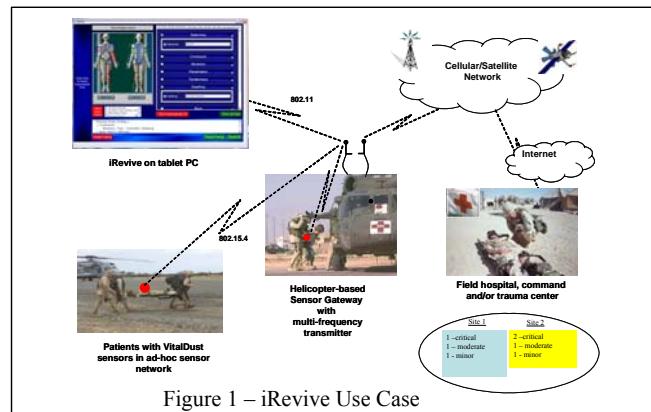


Figure 1 – iRevive Use Case

Figure 1 illustrates how the iRevive application will be used by BMF. The arriving medic places wireless vital sign sensors on one or more patients. Each medic is equipped with a ruggedized tablet PC that captures and displays the real-time sensor data and allows the documentation of observations and treatments. Data capture is automated for vital sign data. Data on observations and treatments is manually entered by the medics. This data entry is guided by a set of rules that enforce consistent and complete capture of data. All of the data captured is subsequently used for billing and other management functions as well as serving as an input for further research. Local medics are linked to the transport aircraft via an 802.11 wireless infrastructure that enables situational awareness of the aircraft crew so they can prepare for any additional medical interventions that may be required. Each transport vehicle is equipped with a base station linking local technicians, command centers, and destination hospitals. This WAN linkage enables global allocation of resources, and increased awareness of the condition of incoming patients at the destination hospital. During patient transport, iRevive continues to capture both sensor data and data recorded by medical personnel. The iRevive application enables the creation and transfer of an electronic patient care record that combines automated capture of vital signs with manually entered data on observations and interventions performed.

1. ACKNOWLEDGMENTS

Our Thanks to the Digital Government Group at NSF

2. REFERENCES

- [1] Hashmi N, Myung D, Gaynor M, Kudlesia V, Tollesen W, Winkler D, Ganesan S, Corwin D, Moulton S. Embedding rules in a mobile database. Proceedings of AICCSA-06 (accepted for publication).
- [2] Baird S, Dawson-Haggerty S, Myung D, Gaynor M, Welsh M, Moulton S. Communicating data from wireless sensor networks using the HL7v3 standard. (Accepted for publication, 2006 IEEE Workshop on Body Sensor Networks).
- [3] Gaynor M, Seltzer M, Moulton S, Freedman J. A Dynamic, Data-Driven, Decision Support System for Emergency Medical Services. NSF workshop on DDDAS, In: Proceedings of the International Conference on Computational Science, June 2005.
- [4] Gaynor M, Myung D, Hashmi N, Kudlesia V, Tollesen W, Winkler D, Ganesan S, Corwin D, Moulton S. Under review, An intelligent pre-hospital patient care system. IJEH, 2007 (International Journal of Electronic Healthcare)
- [5] Gaynor M, Submitted to Conference on Information Systems for Crisis Response and Management, Applications for Emergency Medical Services, NJ, May 2006

Improving the Workflow while Reducing the Costs: Using XML for Web Site Content Management in Government Agencies

J. Ramon Gil-Garcia, Donna Canestraro,
Jim Costello, Andrea Baker, and Derek Werthmuller

Center for Technology in Government
University at Albany, SUNY
187 Wolf Road, Suite 301
Albany, New York 12205
(518) 442-4473

jgil-garcia@ctg.albany.edu

ABSTRACT

XML seems to be a promising alternative for developing and managing e-government web sites. However, there is a lack of empirical research about the use of XML for web site content management in government settings. Based on a multi-method research study, this poster will describe some of the advantages and challenges of using XML. Specifically, it will highlight its potential to streamline the publication workflow while reducing web site maintenance costs. It will also present the potential of XML to achieve greater consistency and allow for multiple formats using a single-source document.

Categories and Subject Descriptors

H.4.2. [Information Systems Applications]: Types of Systems – *e-government applications*.

General Terms

Management, Performance, Economics, Human Factors, Theory.

Keywords

XML, Content Management, Publication Process, Web Sites, E-Government.

1. Description of the Poster

Despite the great promises of the web, managers and leaders are finding their web sites increasingly present challenges of inflexibility, inconsistency, workflow bottlenecks, and new costs [2, 5]. Consequently, government agencies are losing the ability to be responsive and flexible in providing new content or structure. In addition, the costs of maintaining these web sites have become

prohibitive in some cases. Web masters and system administrators have come to realize that the technologies and strategies used in the past to build most web sites are designed to produce individual web pages [3]. However, they do not provide a structure to easily maintain entire web sites, keep them responsive to changing needs, or manage the workflow involved in web content production and maintenance; nor do they facilitate the sharing and reuse of web site content [5].

HTML has several limitations that directly constrain the management of the web sites it is used to produce [3]. The most critical is the one-to-one relationship between file content and web page. With HTML all web pages are individual files, so a site with 1,000 pages is faced with 1,000 individual files to track and maintain. This one to one relationship becomes a problem as sites grow and as multiple distribution channels are considered, for example, multiple browser types [2, 3]. The second key limitation of HTML is a consequence of the embeddedness of content and format. HTML tags contain both content and style information or contain just style information. This embeddedness makes it difficult to manage the workflow of a site. The intermixing of content and style information in an HTML page means that a single source document is difficult if not impossible to create and maintain [2, 5]. In recent years, there have been some advances in HTML such as the use of Cascading Style Sheets (CSS) to separate style from content, representing significant improvements in web site management practices [6]. However, a better solution seems to be needed in order to manage increasingly complex government web sites.

XML is generally understood to be a new technology that supports effective data exchange between applications [1, 4]. However, XML has another value that is much less exploited or understood – it offers an innovative long-term solution to many of the shortcomings of HTML because it structures and describes web content in a meaningful way [3, 5].

In 2001, the Center for Technology in Government (CTG) explored the use of XML to manage an increasing number of publications and documents. The initiative was a response to a workflow problem stemming from the fast growth of CTG's web

site over a five year span. Initially the Web Site had been a place to post simple documents and reports. As CTG grew so did the size and prominence of its web site to a highly complex site of more than 1,300 Web pages, thousands of hyper links, multiple navigation and search routes, interactive applications and ongoing updates. While there were several commercial content management packages available on the market, CTG staff determined that XML's capability to separate content from style to produce multiple formats from one source document would be an effective as well as inexpensive way to solve the workflow problem.

To measure the benefits and challenges associated with the adoption of XML, CTG conducted weekly interviews with staff about the amount of time spent in production and learning the XML format, benefits and drawbacks of the new format. The interview process was conducted to serve as a model for other agencies to conduct their own Return on Investment analysis. From its experience CTG concluded that the initial few months of XML adoption were costly as staff were diverted from routine work to learning the new XML techniques. However, after five months the web team's efficiency increased and the cost of Web Site production dropped rapidly. After approximately 30 weeks CTG's cost on web site production was cut in half. The significant benefits prompted the Center to create a Testbed project to assist state agencies with adopting XML for content management.

Based on semi-structured interviews, a three-wave survey, and analysis of relevant documents, this study analyzes the potential of XML, coupled with business process analysis, as an effective tool for web site content management in government agencies. Five New York State agencies, representing a great diversity in size, goals, and organizational capabilities, are currently participating in the XML Testbed project. The seven-month project began in July 2005 and will end in January 2006. During that time, CTG staff has been helping participants to evaluate the underlying business and workflow processes behind their content management systems. Participants also learned the basics of XML and some transformation languages such as XSL and FO.

Some of the overarching questions guiding this study are:

What are the impacts of using XML for web site content management in government agencies?

Who are the main stakeholders and how they are affected?

What is the return on investment (ROI) associated with the changes in policy, technology, management and data necessary for the use of XML? What are good qualitative and quantitative measures?

What are the success/failure factors of XML adoption and implementation for web site content management in government agencies?

What are some of the effects of XML on the process of developing, managing, and publishing content to the web?

2. Selected References

- [1] Cingil, I., Dogac, A., and Azgin, A., "A Broader Approach to Personalization", *Communications of the ACM*, 43(8), 2000, 136-141.
- [2] Costello, J., Adhya, S., Gil-Garcia, J. R., Pardo, T. A., and Werthmuller, D., *Beyond Data Exchange: XML as a Web Site Workflow and Content Management Technology*. Paper presented at the 2004 Annual Meeting of the Academy of Management: Creating Actionable Knowledge, New Orleans, LA, USA, 2004, August 6-11.
- [3] Hoelzer, S., Schweiger, R. K., Boettcher, H. A., Tafazzoli, A. G., and Dudeck, J., "Value of XML in the Implementation of Clinical Practice Guidelines - The Issue of Content Retrieval and Presentation", *Medical Informatics & The Internet in Medicine*, 26(2), 2001, 131-146.
- [4] Kendall, J. E., and Kendall, K. E., "Information Delivery Systems: An Exploration of Web Pull and Push Technologies", *Communications of the AIS*, 1(Article 14), 1999, 1-43.
- [5] Kerer, C., Kirda, E., Jazayeri, M., and Kurmanowytch, R., *Building and Managing XML/XSL-powered Web Sites: an Experience Report*. Paper presented at the 25th Annual International Computer Software and Applications Conference, Chicago, IL, 2001, October 8-12.
- [6] Lie, H. W., and Saarela, J., "Multipurpose Web Publishing. Using HTML, XML, and CSS", *Communications of the ACM*, 42(10), 1999, 95-101.

Enacting Inter-Organizational E-Government in the Mexican Federal Government

J. Ramon Gil-Garcia
Center for Technology in Government
University at Albany, SUNY
187 Wolf Road, Suite 301
Albany, New York 12205
(518) 442-4473
jgil-garcia@ctg.albany.edu

Luis Felipe Luna-Reyes
Universidad de las Americas
Business School, NE 221J
Santa Catarina Martir
Cholula, Puebla, Mexico 72820
+52 (222) 229-2000 ext. 4536
luisf.luna@udlap.mx

ABSTRACT

This poster will present the results of a research project that investigates inter-organizational e-government in the Mexican federal government. These are e-government initiatives that involve a high degree of collaboration among multiple government agencies and certain degree of information integration. Using the technology enactment framework and a multi-method approach, this study will develop and test models of how interorganizational e-government projects operate. It will explain the complex relationships between organizational structures, institutional arrangements and the selection, design, and use of information technologies. It will also provide some practical lessons for project managers and policy makers interested in e-government.

Categories and Subject Descriptors

H.4.2. [Information Systems Applications]: Type of systems – *e-government applications*.

General Terms

Management, Performance, Standardization, Human Factors, Theory.

Keywords

E-Government, Collaboration, Inter-Organizational, Project Evaluation, Technology Enactment.

1. POSTER DESCRIPTION

The availability of new information and communication technologies (ICTs) and the wave of reform based on principles of the New Public Management have promoted the creation of initiatives that attempt to use technology to enable profound transformations in government [2, 3, 15, 22]. In fact, there are

plenty of examples of governments attempting to transform their governmental structures and improve the quality of the services they provide through the use of ICTs [21]. Consequently, digital government has been considered one powerful strategy for this administrative reform [10, 18]. The promises of digital government range from improved service quality to more effective and efficient government programs and policies [11, 12, 15, 21]. However, digital government initiatives also face a great number of challenges from mismatched data structures and technical incompatibility to organizational resistance and unfavorable institutional arrangements [8, 10, 13, 14].

The promises of digital government can be even more substantial for projects that involved high levels of collaboration and integration among multiple agencies [4, 5]. However, there are also additional challenges for these inter-organizational initiatives [4, 7, 10]. In fact, for some authors these projects represent the highest degree of technological and organizational sophistication [16, 19, 20]. In addition, most of the research about inter-organizational e-government projects has been done in the United States and Europe. The knowledge about potential benefits and challenges in other institutional and social contexts such as Latin America is scarce in the literature.

Important challenges faced by developing countries are associated with the lack of the appropriate technological and human infrastructures, as well as the lack of relevant content in the local language to create a significant social impact. Just to mention a couple of examples, more than 50% of PC computers in Mexican government had a Pentium II processor or older in 2001 [17], and there are only 3.6 million Internet connections for 93.9 million people older than 6 years [1].

This paper will analyze e-government projects that involve high levels of collaboration among multiple government agencies in Mexico. Many of these projects have been promoted and coordinated under the umbrella project called e-Mexico. Within e-Mexico, in the last five years, a team of 13 people has lead efforts to start more than 7,500 Digital Community Centers, to create more than 20 content Internet portals that integrate information and services from government, private companies, and non-governmental organizations, and to develop a technological architecture, which facilitates a deeper integration of information and services. The perceived success of the program is the result of

an innovative approach, involving the creation of content, infrastructure deployment, the development of a technological architecture, and the coordination and collaboration among government agencies.

Therefore, using the technology enactment framework [9, 10] and a multi-method approach, this study will develop and test models of inter-organizational e-government projects. Being a multi-method approach, this study involves semi-structured interviews, three case studies, and a survey [6, 23]. First, semi-structured interviews will be conducted to project managers of 15-20 interorganizational e-government projects managed by the Mexican federal government. Then, based on the results of the initial interviews, three initiatives will be selected and case studies will be prepared for these three projects. The case studies will involve additional semi-structured interviews, documentation analysis, and an evaluation of the web portal of the initiative. After the initial interviews and in parallel to the case studies, a survey to all the participating organizations in the 15-20 interorganizational initiatives will be conducted. The objective of this complex design is to understand some of the mechanisms and results of inter-organizational e-government in Latin American contexts. It will also provide some evidence of the similarities and differences between Latin America and other contexts.

ACKNOWLEDGMENTS

The research reported here is supported by the Consejo Nacional de Ciencia y Tecnología (CONACYT-Mexico) grant SEP-2004-C01-46507. The views and conclusions expressed in this paper are those of the authors alone and do not necessarily reflect the views or policies of CONACYT.

2. References

- [1] AMIPCI, *Estudio AMIPCI de Internet en México 2005*, 2005, Retrieved October, 2005, 2005, from http://www.amipci.org.mx/docs/Presentacion_Estudio_AMIPCI_2005_Presentada.pdf
- [2] Arellano-Gault, D., "Challenges for the New Public Management. Organizational Culture and the Administrative Modernization Program in Mexico City (1995-1997)", *American Review of Public Administration*, 30(4), 2000, 400-413.
- [3] Barzelay, M., *Breaking through Bureaucracy*, University of California Press, Berkeley and Los Angeles, CA, 1992.
- [4] Caffrey, L., *Information Sharing Between & Within Governments*, Commonwealth Secretariat, London, 1998.
- [5] Cook, M., Dawes, S. S., Juraga, D., Werthmuller, D. R., Pagano, C. M., and Schwartz, B. F., *Bridging the Enterprise: Lessons from the New York State-Local Internet Gateway Prototype*, Center for Technology in Government, University at Albany, SUNY, Albany, NY, 2004.
- [6] Creswell, J. W., *Research Design. Qualitative, Quantitative, and Mixed Methods Approaches*, SAGE Publications, Thousand Oaks, CA, 2003.
- [7] Dawes, S. S., and Pardo, T. A., Building Collaborative Digital Government Systems. Systematic Constraints and Effective Practices. In W. J. McIver & A. K. Elmagarmid (Eds.), *Advances in Digital Government. Technology, Human Factors, and Policy* (pp. 259-273). Kluwer Academic Publishers, Norwell, MA, 2002.
- [8] Fletcher, P. D., Policy and Portals. In W. J. McIver & A. K. Elmagarmid (Eds.), *Advances in Digital Government: Technology, Human Factors, and Policy*. Kluwer Academic Press, Norwell, MA, 2002.
- [9] Fountain, J. E., *Enacting Technology: An Institutional Perspective*, John F. Kennedy School of Government, Harvard University, Cambridge, MA, 1995.
- [10] Fountain, J. E., *Building the Virtual State. Information Technology and Institutional Change*, Brookings Institution Press, Washington, D.C., 2001.
- [11] Gant, J. P., and Gant, D. B., *Web portal functionality and State government E-service*. Paper presented at the 35th Hawaii International Conference on System Sciences, Hawaii, 2002
- [12] Garson, G. D., The Promise of Digital Government. In A. Pavlichev & G. D. Garson (Eds.), *Digital Government: Principles and Best Practices* (pp. 2-15). Idea Group Publishing, Hershey, PA, 2004.
- [13] Gil-García, J. R., and Pardo, T. A., "E-Government Success Factors: Mapping Practical Tools to Theoretical Foundations", *Government Information Quarterly*, 22(2), 2005, 187-216.
- [14] Glassey, O., "Developing a one-stop government data model", *Government Information Quarterly*, 21(2), 2004, 156-169.
- [15] Heeks, R., *Reinventing Government in the Information Age. International Practice in IT-Enabled Public Sector Reform*, Routledge, New York, 1999.
- [16] Holden, S. H., Norris, D. F., and Fletcher, P. D., "Electronic Government at the Local Level: Progress to Date and Future Issues", *Public Performance and Management Review*, 26(4), 2003, 325-344.
- [17] INEGI, *Estructura Porcentual de las Computadoras Personales de la Administración Pública por tipo de Procesador por cada nivel de la Administración Pública*, 2001, Retrieved October 2005, 2005, from <http://www.inegi.gob.mx/est/contenidos/espanol/rutinas/ept.asp?t=tinf003&c=3425>
- [18] Kramer, K. L., and King, J. L. (2003). *Information Technology and Administrative Reform: Will the Time After E-Government Be Different?* Unpublished manuscript, Irvine, CA.
- [19] Layne, K., and Lee, J., "Developing fully functional E-government: A four stage model", *Government Information Quarterly*, 18, 2001, 122-136.
- [20] Moon, M. J., "The Evolution of E-Government Among Municipalities: Rhetoric or Reality?" *Public Administration Review*, 62(4), 2002, 424-433.
- [21] OECD, *The e-Government Imperative*, Organisation for Economic Co-operation and Development, Paris, France, 2003.
- [22] Osborne, D., and Gaebler, T., *Reinventing Government. How the Entrepreneurial Spirit is Transforming the Public Sector*, Plume, New York City, 1992.
- [23] Yin, R. K., *Case Study Research. Design and Methods*, Sage Publications, Thousand Oaks, CA, 2003.

Estimating Freight Flows for Metropolitan Highway Networks Using Secondary Data Sources

Genevieve Giuliano

University of Southern California
School of Policy, Planning, and
Development
Ralph and Goldy Lewis Hall 216
Phone: (+1) 213-740-3956
Email: giuliano@usc.edu

Peter Gordon

University of Southern California
School of Policy, Planning, and
Development
Ralph and Goldy Lewis Hall 321
Phone: (+1) 213-740-1467
Email: pgordon@usc.edu

Qisheng Pan

Texas Southern University
School of Public Affairs
HH 334E
Phone: (+1) 713-313-7221
Email: pan_qs@tsu.edu

ABSTRACT

In this paper, we suggest that it is possible to estimate most of a metropolitan area's highway network truck shipments from secondary data sources, using these sources to generate relatively inexpensive and updateable link-specific estimates.

Categories and Subject Descriptors

E.4 [Coding and Information Theory]: Data compaction and compression
I.6 [Simulation and Modeling]: Model Development
E.0 [General]: Data

General Terms

Measurement, Design, Experimentation, Theory

Keywords

Freight Flow, Secondary Data Sources, Highway Network.

1. INTRODUCTION

Trade and shipments between countries, regions and cities have been expanding dramatically. As a result, freight traffic is an increasingly important factor in urban transportation systems. Transportation planners and social scientists studying changing patterns of trade have mostly relied on rules-of-thumb and/or infrequent and expensive shipper surveys to estimate metropolitan level freight flows. These two approaches are no longer adequate. We suggest that it is possible to estimate most of a metropolitan area's highway network truck shipments from secondary data sources, using these sources to generate relatively inexpensive and updateable link-specific estimates. This approach is made possible by advances in web services and computational workflows. Using new information management

and processing techniques allows for efficient access and processing of data across many disparate sources.

The major research steps in generating link specific freight flows are the following:

- (1). Utilize a regional input-output transactions table to estimate intraregional commodity-specific trip attractions and trip productions, and allocate these to small-area units.
- (2). Estimate commodity-specific interregional and international trip attractions and trip productions for those locations where airports, seaports, rail yards or regional highway entry-exit points are located.
- (3). Create a regional commodity origin-destination matrix using estimates from (1) and (2).
- (4). Load the O-D matrix onto a regional highway network with known passenger flows.

This document discusses the first two steps. It is complicated by the fact that freight data from the most important data sources are described via various (often independent) classificatory systems and definitions. Much of our work has been devoted to reconciling data from these various sources.

Here, we describe a prototypical application of our approach to the Los Angeles metropolitan area (the five-county CMSA). We include the steps required to reconcile several commodity flow data sources. A major challenge is the lack of a common industry sector classification system. We have created a set of industrial sectors, the "USC Sectors," that are a basis for our approach.

2. DATA SOURCES AND RECONCILIATION

Our research objective is to develop a method for estimating metropolitan freight flows that is easily updated and is transferable across metropolitan areas. Our data sources are: 1) a regional input-output data file from the Minnesota IMPLAN Group; 2) the Commodity Flow Survey; 3) WISERTrade airport and seaport data; 4) the California Department of Transportation's ITMS file; 5) jobs by transportation analysis zone (TAZ); 6) Waterborne Commerce

of the United States; and 7) data supplied by major airports in the region and complementary data from RAND.

Our approach requires the use of commodity- or sector-specific data from a variety of sources. This made it necessary for us to reconcile various classifications; we developed the 47-sector system of "USC Sectors" for this purpose.

3. INTRAREGIONAL SHIPMENTS

To estimate truck shipments within a region, we first estimate commodity supply and demand within the region to and from local enterprises. We generate a set of O-D matrices that describe commodity flows in and out of small areas for 9 aggregated SCTG commodity sectors (number of sectors is determined by limitations of inter-regional data sources).

Estimating the attractions and productions of commodities for each TAZ requires a regional input-output transactions table and small area employment data (see Cho et al., 1999). The regional input-output model, with out-of-region shipments removed, provides the basis for estimating local supply and demand generated zone-to-zone shipments, once the regional coefficients are combined with small-area jobs-by sector data. The IMPLAN input-output transactions table provides the dollar values of inter-sector commodity flows that serve household consumption and the parts of final demand not associated with households.

There are two major estimation equations:

$$D_i^z = \sum_j a_{ij} X_j^z \quad (1)$$

where, X_j^z = total regional output of commodity j in zone z , given base year employment in sector j and zone z ,

a_{ij} = i, j th element of A, the matrix of value demand coefficients for the (open) input-output model, representing the flow from i to j per unit output of j .

D_i^z = freight flows attracted from sector i in response to demand in zone z . D_i^z includes shipments of commodity i to zone z from transshipment zones (imports) and other zones.

Similarly, Equation (2) calculates the total supply of output j provided by zone z ,

$$O_j^z = \sum_i b_{ij} X_i^z \quad (2)$$

where, X_i^z = total regional output of commodity i in zone z , given base year employment in sector i and zone z ,

b_{ij} = i, j th element of B, the matrix of value supply coefficients for the (open) input-output model, representing flow from i to j per unit output of i .

O_j^z = freight flows produced in zone z to satisfy the demands by sector j . O_j^z includes the shipments of commodity j to transshipment zones to other zones.

4. INTERREGIONAL AND INTERNATIONAL SHIPMENTS

The Commodity Flow Survey collects data on the movement of goods within the U.S. every five years. The CFS "inbound" and "outbound" flows include all flows that are shipped to and from local area establishments, even if they are shipped to or from a transshipment point for purposes of international and interregional trade; at the metropolitan level, CFS modal and sectoral data are limited, especially for inbound commodity flows. We therefore rely on some of the California, West-region and Pacific-division CFS data.

IMPLAN provides shipments data which can be reconciled with the CFS data in terms of the definitions of freight flows. The bridge table for IMPLAN and CFS sectoral classifications makes the aggregation of IMPLAN sectors into SCTG sectors possible. In general, it is the joint use of CFS and IMPLAN data that makes the estimation of interregional and international shipments by mode and by sector complete.

Spatial allocation of freight trip-ends is also conducted. Interregional and international shipments to and from the Los Angeles Consolidated Metropolitan Statistical Area (CMSA) pass through the region's two seaports, five airports, three rail yards and six major highway entry/exit points. The spatial allocation of inbound and outbound truck and rail shipments are based on data from 1996 ITMS report from the California Department of Transportation. For the spatial allocation of airborne and waterborne commodities, freight shares for five airports were based on the airport statistics and seaports data from WCUS or WISERTrade (see Giuliano et al, 2005 for details).

The result of this series of computations is an origin-destination matrix of freight flows in dollar units, by mode. Highway flows are converted to vehicle equivalent units and then assigned to the highway network.

ACKNOWLEDGMENTS

This research is supported by NSF grant 0138998.

ADDITIONAL AUTHORS

Jiyoung Park (University of Southern California, email: jiyoungp@usc.edu), Lanlan Wang (University of Southern California, email: lanlanwa@usc.edu)

REFERENCES

- [1] Cho, S-B. et. al.. Integrating Transportation Network and Regional Economic Models to Estimate the Costs of a Large Earthquake: NSF Draft Report (1999).
- [2] Giuliano, G. et. al.. Estimating Freight Flows for Metropolitan Area Highway Networks Using Secondary Data Sources, working paper (Jul. 2005).
- [3] Gordon, P., Pan, Q. Sh.. Assembling and Processing Freight Shipment Data: Developing a GIS-Based Origin-Destination Matrix for Southern California Freight Flows, METRANS report (Jun. 2001).

TIME-CRITICAL INFORMATION SERVICES

Update on Exploratory Analysis of Emergency Response and Related E-Governmental Services

Thomas A. Horan, Ph.D.

Claremont Graduate University
School of Info. Systems & Tech.
130 East Ninth, Claremont, CA
Tom.Horan@cgu.edu

Michael Marich

Claremont Graduate University
School of Info. Systems & Tech.
130 East Ninth, Claremont, CA
Michael.Marich@cgu.edu

Ben Schooley

Claremont Graduate University
School of Info. Systems & Tech.
130 East Ninth, Claremont, CA
Ben.Schooley@cgu.edu

ABSTRACT

Time-critical information services (TCIS) is defined as the medical necessity to deliver emergency services as rapidly as possible, coupled with the dependence of these services upon accurate and timely information from multiple organizations. This paper provides a discussion of the authors' current National Science Foundation (NSF)-funded project involving TCIS, with specific reference to its use in emergency response and related e-government services. The project includes the development of a general framework for understanding and researching the end-to-end performance of inter-organizational e-governmental services and findings from an expert workshop held at the National Center for Digital Government. The TCIS invitational workshop allowed for expert (academic and practitioner) input and feedback on the TCIS dimensions and the best means for understanding their occurrence in on-the-ground emergency medical services (EMS). Workshop participants analyzed TCIS from a socio-technical perspective and provided conceptual, practitioner and methodological critiques and suggestions. Overall, participants found the concept of TCIS to be a valid model for understanding, researching, and developing e-government systems within the specific context of emergency response as well as within the broader context of time-critical services to the public. Workshop recommendations focused on the need to closely assess inter-agency and inter-organizational *information exchanges* along and between three levels: technical, organizational, and governance.

1. INTRODUCTION

This paper provides the highlights of the authors' research in the area of TCIS, with specific reference to its use in emergency response and related e-governmental services. The paper is composed of three parts: (1) a description of the current project activities, (2) a summary of the published and unpublished research contributions, and (3) a discussion of the successes, challenges, and plans for the coming year.

2. CURRENT PROJECT ACTIVITIES

For the last three years, the authors have examined TCIS within the context of rural emergency response [2]. Based on prior research work and case studies, we have determined that (1) EMS is a fundamentally inter-organizational system and (2) there are a range of technical, organizational, and governance factors that

affect system operations and performance. It also became clear that there is a need for greater conceptualization related to inter-organizational, end-to-end performance from a time-critical perspective – both for EMS specifically and for e-government generally. The NSF Digital Government SGER Grant (Award no. 0508938) provided an opportunity for such an exploration.

One of the first research tasks was to construct a conceptual model, drawing upon a synthesis of findings from prior research, as well as findings from literature review. Based on this preliminary analysis, there are several components that would enter into a conceptual model for time-critical information services, both in regard to EMS specifically and other public services generally. These components include: 1) the time and information critical elements of the service, 2) inter-organizational linkages that include both qualitative organizational elements as well as "hard" information flow elements, 3) end-to-end elements that consider performance metrics within and across the process flow, and 4) context variation elements such as normal versus peak conditions (in terms of service demand).

The team then presented this model at an expert research workshop for further refinement. The workshop featured experienced academics, researchers, and practitioners and was conducted in April 2005 at Harvard University in cooperation with the National Center for Digital Government. Additional co-sponsorship was provided by the State and Local Policy Program, Humphrey Institute, University of Minnesota in collaboration with the ITS Institute, University of Minnesota.

The workshop was organized in four sections: concept review, methodological discussion, case applications, and future directions. During the first session, participants discussed the system dimensions of the TCIS model, with specific attention on the governance of TCIS. During the methodological session, participants highlighted the challenge of identifying, defining and/or using appropriate performance metrics, as well as the need to include the societal impacts, such as lives saved, as a metric. Case applications included discussion of on the ground deployments occurring in California, Minnesota, and New England and revealed the need to look at technical, organizational and governance dimensions simultaneously. The future directions session highlighted the opportunity to advance TCIS concepts in

new “next generation 911” activities as well as future research and outreach activities.

In sum, the NSF-sponsored workshop highlighted the need to understand multiple nested dimensions of interactions that occur in time-critical events. Participants found the concept of TCIS to be a valid model for understanding, researching, and developing e-government systems within the specific context of emergency response as well as within the broader context of time-critical services to the public. The workshop helped researchers obtain additional insight into how EMS collaboration can vary as a function of technical, organizational, and governance systems. It also gave the research team confidence that various sites selected for future field study work were appropriate venues to examine collaborative systems in action.

3. RESEARCH CONTRIBUTIONS

The major research contribution of this project is to highlight the importance of the ‘end-to-end’ nature of performance and the critical role of information sharing across organizations to maximize performance results. For example, it makes little difference for a 9-1-1 operator to dispatch quickly if the ambulance takes a very long time to arrive and/or goes to the wrong location. Measuring effectiveness across organizations (end-to-end performance) is essential to understanding how public services are delivered to the public, the level of service (timeliness, quality) with which they are delivered, and how the network can be improved to deliver better services in an information-critical and time-critical manner.

The challenge is how to implement this “end-to-end” concept within and across emergency provider organizations. In our research, we have found that service hand-offs were incrementally improving the time and information flows. However, the missing element was an integrated “organizational awareness” regarding interoperation between organizations [1]. While information systems are often implemented to address separate silos of a governmental process, the end-to-end nature of TCIS facilitates or at least allows for information systems that can report on overall system performance.

The concept of TCIS and ‘end-to-end’ performance has been reported in a variety of research and publication domains. Project results have been integrated into the University of Minnesota’s Transportation Center research program, State and Local Policy program, the NSF-sponsored digital government program, and a number of other research follow-on activities are underway (see below). Findings have been presented and published at a variety of research forums. For example, a summary article has been accepted for publication in the *Communications of the ACM* [2], and other research articles are in preparation.

4. SUCCESSES, CHALLENGES, PLANS

Subsequent to the research symposium, the authors have had additional discussions and interactions with the participants including responding to a Request for Information (RFI) from the National Highway Traffic Safety Administration (NHTSA) on Next Generation 9-1-1 Systems. These interactions have further

confirmed the value of continuing research efforts to expand the TCIS model. As such, our future research falls into two primary areas: (1) understanding the needs of the users of TCIS systems from multiple dimensions and (2) introducing the notion of performance into the national systems architecture for EMS systems.

The TCIS symposium highlighted the need to understand the context of inter-organizational information exchange from three levels: technological, organizational, and policy. Thus, the authors plan to examine information hand-offs from one emergency response organization to another across different deployments in the U.S. through a grounded case study approach. The goal would be to understand the operational system, and then to understand information hand-offs from technological, organizational, and policy perspectives. Such context will provide a way to understand how performance information is acquired and shared across organizations, the barriers and synergies to sharing such information, and the requirements for overcoming barriers to including performance metrics.

Within the U.S., work is currently underway within the Department of Transportation (DOT) at a national level to develop the next generation 9-1-1 (NG 9-1-1) system that integrates voice, video, and data. However, the present research by the authors has shown that there exists a challenge to ensure that adequate information related to the end-to-end system performance is captured. From this perspective, the authors plan to develop a framework for incorporating performance information into the Intelligent Transportation Systems (ITS) architecture as well as the NG 9-1-1 systems architecture and preliminary concept of operations (ConOps) proposed by the U.S. DOT for EMS. The authors intend to provide a set of recommendations for incorporation into these systems.

5. ACKNOWLEDGEMENTS

The TCIS Workshop, sponsored by the Digital Government Program, National Science Foundation (Award no. 0508938) was conducted in cooperation with the National Center for Digital Government, Harvard University. Additional co-sponsorship was provided by the State and Local Policy Program, Humphrey Institute, University of Minnesota in collaboration with the ITS Institute, University of Minnesota.

6. REFERENCES

- [1] Horan, T., and Schooley, B. (2005a). Time-Critical Information Services. *Communications of the ACM*. (Accepted and forthcoming).
- [2] Horan, T. and Schooley, B. (2005b). Interorganizational Emergency Medical Services: Case Study of Rural Wireless Deployment and Management. *Information Systems Frontiers*. 7(2), pp. 155-173.

Entity Consolidation and Alignment in Semi-Structured Data Sources

Eduard Hovy, Andrew Philpot and Patrick Pantel

Information Sciences Institute
University of Southern California
4676 Admiralty Way
Marina del Rey, CA 90292

{hovy,philpot,pantel}@isi.edu

ABSTRACT

A large portion of collected data is stored in semi-structured form, i.e., organized or related through columns or lists but without any formal schema or metadata. For example, many government organizations and companies feature employee directories and project affiliations in HTML tables on their websites. Making sense of these requires automatic methods for data alignment, matching and/or merging. Here, we describe *Guspin*, a tool for automatically consolidating entities and for aligning data across semi-structured data sources. Our project, based on principles of information theory, measures the relative importance of data, leveraging them to quantify the similarity between entities. We have applied our technology to discover duplicates and perform alignments for data sources provided by the Environmental Protection Agency and California environmental agencies.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Filtering.

General Terms

Algorithms, Experimentation.

Keywords

Information theory, mutual information, semi-structured data, equivalence class detection, entity consolidation.

1. *Guspin*¹: A Data Modeling Portal

Guspin is a general purpose tool for finding equivalence classes, consolidating entities, and aligning data within a population of semi-structured data. It provides a simple user interface where a user uploads one or multiple data files containing observations for a population. The system inspects the data, identifies patterns, suggests alignments and consolidations, provides a browsing interface for viewing the analysis, and permits downloading of the analysis for further examination.

2. Case Study A: Entity Consolidation

The Environmental Protection Agency (EPA) maintains a centrally managed database, called the Facilities Registry System (FRS), recording American facilities subject to environmental regulations (e.g., refineries, gas stations, manufacturing sites, etc.) Duplicates exist in the FRS since it is compiled from various local and state jurisdictions, which often have different ways of representing data. Our goal on this data set is to automatically discover the duplicate entries.

We obtained from the EPA a sample of the FRS. Each record includes the address, state, zip code, facility name, etc. for a particular facility. Through *Guspin*'s web interface, we upload the FRS data and then *Guspin* measures the mutual information between entities and observations (e.g., address, emission statistics, codes, etc.), computes the similarity between each pair of entities, and clusters entities into equivalence classes. One can search for individual entities by using *Guspin*'s search feature. For example, *Guspin* discovered that facility 189 consolidates with facilities 300 and 79. Figure 1 shows the results of launching a search for facility 189's most similar entities (i.e., potential duplicates). For each similar entity, the similarity score is shown along with a "why?" link, which enables the user to compare the observations of the two facilities (important observations are used to compute the similarity between entities).

Figure 2 illustrates two such comparisons: a) a comparison between the observations for facilities 189 and 79; and b) a comparison between the observations for facilities 189 and 300. Observations colored in blue and in green were observed for only one of the two facilities. Red observations, however, were shared by both facilities. Figure 2 lists observations in descending order of mutual information scores. For very similar entities, we therefore expect that most important observations (those at the top of the list) will be colored red. In fact, note that even though Figure 2 shows that facilities 189 and 79 share fewer common observations than facilities 79 and 300, the similarity between facilities 189 and 79 is greater since more *important* features are shared (i.e., they have more red features at the top of the list).

3. Case Study B: Data Alignment

Continuing our relationship with the California Air Resources Board (CARB) and various California Air Quality Management Districts (AQMDs), built during our DG project, we obtained emissions inventories (a comprehensive description of emitters and emission statistics) submitted annually by California AQMDs to CARB. We applied *Guspin* to the inventories to automatically

¹ *Guspin* is available from <http://guspin.isi.edu>.

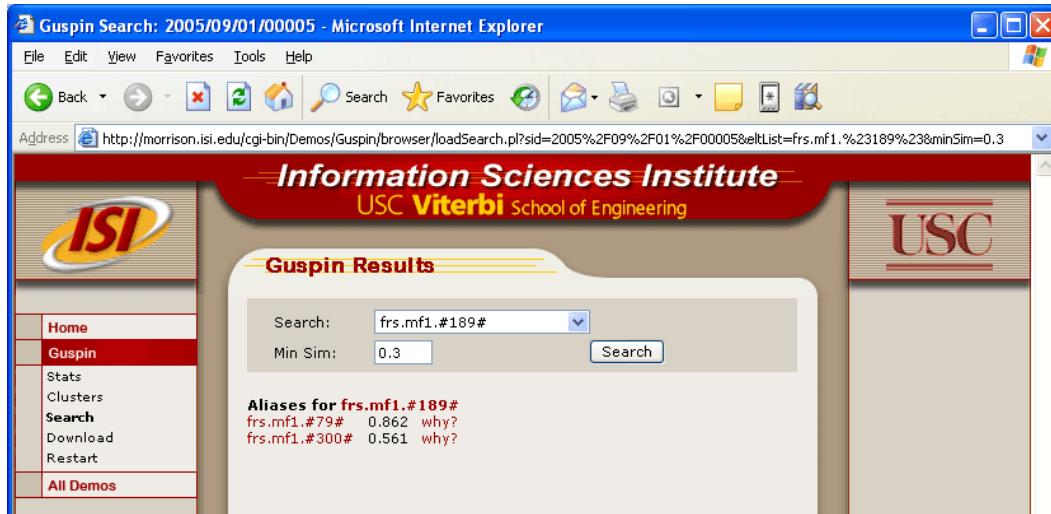


Figure 1. *Guspин*'s search interface for displaying an entity's most similar entities. In this example, we see that facility 189 from EPA's Facilities Registry System is most similar to facilities 79 and 300. Clicking on a facility displays its observations. Clicking on "why?" compares the observation data from facility 189 with those from facilities 79 and 300.

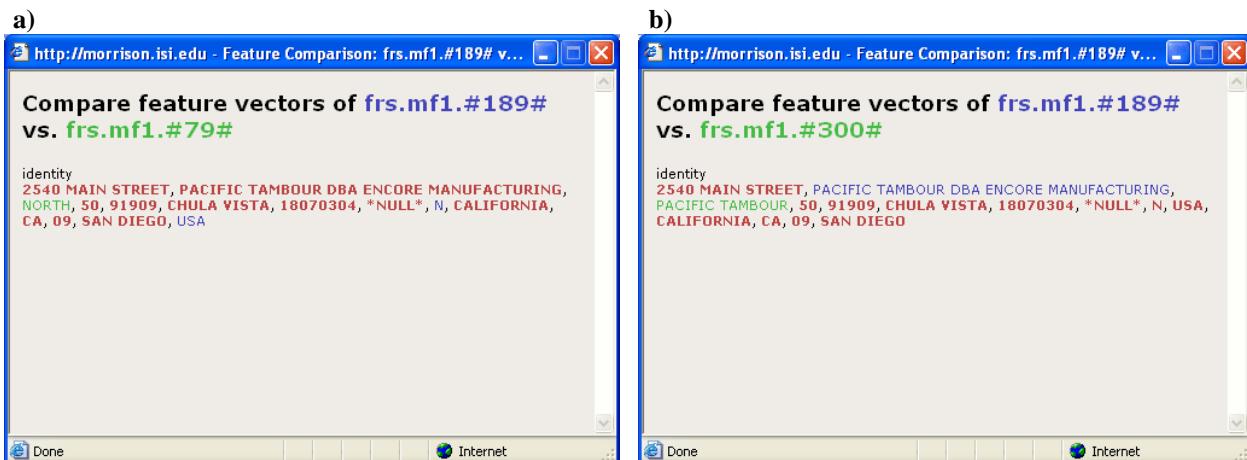


Figure 2. *Guspин* comparison of two entities' observations: a) comparison of the observations for facilities 189 and 79 (similarity = 0.862); b) comparison of the observations for facilities 189 and 300 (similarity = 0.561). Observations are sorted in decreasing order of pointwise mutual information scores. Observations colored in blue and green are shared by only one of two facilities, whereas red observations are shared by both.

discover overlapping entries across data sources. Below, we summarize *Guspин*'s performance on the CARB and Santa Barbara County Air Pollution Control District 2001 emissions inventories:

- with 100% accuracy, *Guspин* extracted 50% of the matching facilities;
- with 90% accuracy, *Guspин* extracted 75% of the matching facilities;
- for a given facility and the top-5 mappings returned by *Guspин*, with 92% accuracy, *Guspин* extracted 89% of the matching facilities.

4. Conclusions

Researchers, organizations and government agencies working with semi-structured data are in need of a tool for discovering

duplicate or overlapping data. Our project, based on principles of information theory, measures the importance of observations and then leverages these to quantify the similarity between entities. Though the technology is applicable to a wide range of applications, we have built *Guspин* to address the general problems of consolidating entities and aligning data across semi-structured data sources. *Guspин* has been applied to solve real problems faced by the Environmental Protection Agency.

Guspин may be applied to several other tasks. For example, it can be used to identify occurrences of plagiarism in essays represented by the words they contain or it can be used to find co-regulated genes represented by their expressions in a series of micro-array experiments. It is our intention to promote *Guspин* to all government agencies as a portal for helping in the collection, disambiguation, and analysis of their data.

UNeGov.net – Community of Practice for Electronic Governance

Tomasz Janowski, Elsa Estevez, Irshad Khan, Adegboyega Ojo

United Nations University, International Institute for Software Technology (UNU-IIST)
P.O. Box 3058, Macau
+853 5040443
{tj, elsa, ik, ao}@iist.unu.edu

ABSTRACT

The paper presents an initiative by UNU-IIST to build a global Community of Practice interested in developing, sharing and applying concrete solutions for e-Governance – UNeGov.net. We present the rationale for the initiative, along with its mission, objectives and activities. We also describe a novel approach to collaborative problem solving supported by the UNeGov.net Portal. The approach is based on a repository of resources relevant to e-Governance to underpin a process of formulating, exploring, matching and refining abstract problem descriptions into concrete solutions, enriching the repository in the process.

Categories and Subject Descriptors

J.1 [Computer Applications]: Communities of Practice

General Terms

Algorithms, Experimentation, Human Factors, Standardization

Keywords

Community of Practice, Electronic Governance, Cooperative Problem Solving, Knowledge Sharing

1. INTRODUCTION

Governments worldwide are under pressure to address public needs, to support local industries, to deliver high-quality public services, etc. In response, they engage in Public Sector Reform and develop e-Governance - leveraging the use of Information and Communication Technology (ICT) to bring about customer orientation, businesslike management and other public sector reforms [3]. In doing so, they face many challenges [2], e.g. how best to: lead organizational changes despite resistance from civil servants; establish cross-agency projects against hierarchical government structures; build long-lasting technology solutions, while facing technology volatility; rely on the private sector to deliver public services, while avoiding vendor lock-in strategies.

2. INITIATIVE

Alongside the challenges facing public managers responsible for technology and reform initiatives, there is a growing experience on how such challenges can be addressed. UNeGov.net is a forum to share such experiences and develop localized solutions.

The mission of UNeGov.net is to build a Community of Practice [1] focused on developing, sharing and applying concrete solutions for e-Governance through research, development and community collaboration, with emphasis on developing countries.

In line with its mission, UNeGov.net has the following objectives: (1) advance the practice of e-Governance, (2) focus on solutions to concrete problems, (3) build consensus on best practices, (4) consider the challenges facing Developing Countries, (5) facilitate the sharing of experiences and resources and (6) support interactions between practitioners and experts.

UNeGov.net engages in nine kinds of activities:

- 1) *Portal* - A repository of resources for common use across the Community, with support for collaborative problem-solving.
- 2) *Workshops* - Share experiences, identify concrete issues of interest to governments, discuss reusable solutions supported by research and cooperation, and build a Community.
- 3) *Schools* - Organize schools and courses for public IT managers, CIOs and industry leaders, particularly from Developing Countries, on various aspects of e-Governance.
- 4) *Projects* - Jointly apply for funding and execute projects through community-wide cooperation to advance the state of e-Governance in particular countries and globally.
- 5) *Reports* - Document experience of individual countries in Electronic Governance through Country Reports. Document the state-of-the-art globally through Thematic Reports.
- 6) *Surveys* - Carry out a survey on the global state of Electronic Governance through member contributions on individual countries and analysis of existing survey series.
- 7) *Curriculum* - Develop a curriculum for public managers and CIOs and help members adapt, customize and implement it according to the needs of each country.
- 8) *Conference* - Establish a global forum for researchers, practitioners and developers to present the latest findings on the theory and practice of Electronic Governance.
- 9) *Practice* - Create a framework for collaborative problem-solving, complementing the systematic solution-building process with deep learning experience by members.

Figure 1 shows how such activities are mapped into objectives.

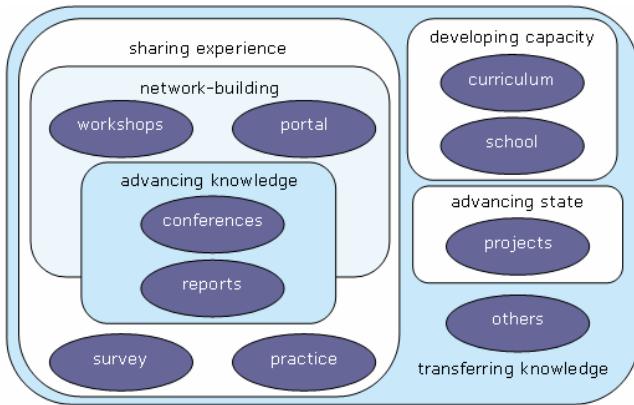


Figure 1. Activities vis-à-vis objectives of UNeGov.net

3. COMMUNITY

The UNeGov.net Community is the centre of all activities – members initiate and execute such activities, resulting in more experience and content, and enlargement of membership.

The Community consists of experts in different fields related to e-Governance: Public Administration, Information Technology, Knowledge Management, Political Sciences, etc. It also comprises practitioners from governments, industry and academia: public managers responsible for technology and reform initiatives, decision makers, CIOs, government service providers, etc.

Figure 2 presents a matrix mapping various relevant areas of specialty (for experts) and affiliations (for practitioners) against different countries the members come from. On this basis, the figure also illustrates the idea of thematic versus country reports.

		Countries:								e-Governance in Argentina, Country Report		
		Tunisia	Palestine	Vietnam	Nepal	India	Cameroun	Nigeria	China	e-Governance in Argentina, Country Report		
Experts	Themes:	legislation										
	financing											
	organization											
	Change Management for Public Administrations, Thematic Report											
	electronic democracy											
	electronic administration											
	government											
Practitioners	industry											
	academia											
	civil society											

Figure 2. UNeGov.net Community matrix

4. COOPERATIVE PROBLEM SOLVING

The practice activity of UNeGov.net applies a novel technical approach to Cooperative Problem Solving in Virtual Communities of Practice. The approach defines a systematic process of solution-building for a given problem description. The process relies on a repository of various web resources – papers, projects, software, people, problems, solutions, etc.; properties – data about or relationships between resources; and statements – triples of subject (resource), property and object (resource or data) [4].

Problem solving is carried out in six stages:

- 1) *problem description* – A member describes a problem from its own practice and adds it to the repository as a resource.
- 2) *problem exploration* – By exploring the problem, relevant resources, properties and statements are gradually added.
- 3) *problem matching* – The problem is matched against similar problems, solved or unsolved, described in the repository.
- 4) *solution design* – The first solution is proposed for the problem, including its decomposition into sub-problems, each added to the repository as a standalone problem description.
- 5) *solution refinement* – The solution is refined by adding relevant resources, properties and statements, and integrating sub-problem solutions into solution, once available.
- 6) *solution deployment* – When all sub-problems are solved, add a statement relating abstract problem with concrete solution.

The process can be carried out fully collaboratively since sub-problems can be assigned to different members, who in turn apply the same process for solving them. Figure 3 illustrates the process.

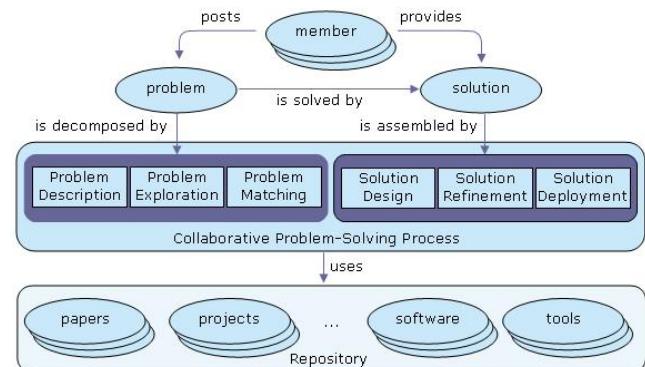


Figure 3: UNeGov.net and Cooperative Problem Solving

5. CONCLUSIONS

UNeGov.net is a newly established initiative of UNU-IIST to build a Community of Practice for e-Governance. This paper highlights the mission, objectives and activities of UNeGov.net, its community and the novel approach to cooperative problem-solving. More information can be found at www.unegov.net.

6. ACKNOWLEDGEMENTS

We wish to thank Clara Casalini, Brian Iu and Gabriel Oteniya for support and collaboration on establishing UNeGov.net.

7. REFERENCES

- [1] Cambridge, D., Kaplan, S. and Suter, V. *Community of Practice Design Guide*, 2005.
- [2] Heeks, R., *Causes of e-Government Success and Failure: Factor Model*, University of Manchester, 2003.
- [3] OECD - Organization for Economic Cooperation and Development, *The e-Government Imperative*, 2003.
- [4] W3C. Semantic Web Activity. *Resource Description Framework*, 2005. <http://www.w3.org/RDF/>

Unraveling Shared Services using Simulation

Marijn Janssen

Faculty of Technology, Policy and Management,
Delft University of Technology
Jaffalaan 5, NL-2628 BX, Delft,
The Netherlands,
Tel. +31 (15) 278 1140,
m.f.w.h.a.janssen@tudelft.nl

René W. Wagenaar

Faculty of Technology, Policy and Management,
Delft University of Technology,
Jaffalaan 5, NL-2628 BX, Delft,
The Netherlands,
Tel. +31 (15) 2788077
r.w.wagenaar@tudelft.nl

ABSTRACT

Existing and new services can potentially be shared among governmental organizations to achieve economies of scale and scope. In designing shared services in public administration various views need to be integrated in order to make well-sounded trade-offs. In this respect, it is often argued that there is insufficient support for structural reorientation and transformation of public organizations.

The overall objective of this ongoing research is to support the planned change and transformation of public administration through the implementation and deployment of shared services. In this poster presentation we will demonstrate how flow-oriented, discrete-event simulation can be used to model and unravel shared services by capturing the organizational, business process, and application views. In this way simulation can support the diverse stakeholders that are involved in the structural reorientation of public administration. Moreover, we will show using case study how simulation can support decision-making and facilitate the making of trade-offs through what-if analyses.

Categories and Subject Descriptors

H.1.1. [General System Theory]: Models and Principles – *Systems and Information Theory*; J.1 [Administrative Data Processing]: Administrative Data Processing – *Government*; K.6.4 [Management of computing and Information Systems] Management of Computing and Information Systems– *System Management*

General Terms

Management, Performance, Design, Economics.

Keywords

Transformation, Simulation, Animation, Decision support, Shared services

1. INTRODUCTION

Approaches characterized by the focus on business processes can contribute to the management of changes in public organizations

[7]. Beynon-Davies and Williams [2] found that in the UK there was not enough emphasis on the engineering of both business processes and information systems. In attempts to re-design public management operations through the extensive use of information technology, it appears that there is still insufficient support for business process re-engineering and structural reorientation and transformation in public organizations that are characterized by a multitude of stakeholders involved in the decision making. [3],[6]. As such there is a need for tools facilitating the transformation of public organizations.

Discrete-event simulation can be used to experiment with a system prior to implementation and animation can be used to visualize the behavior of the system under study. As a technique, simulation is one of the most widely used in operations research and management science [8]. In this poster presentation we will demonstrate how simulation can be used to unravel the organization, business processes and information systems involved in the delivery and deployment of shared services and evaluate various shared service arrangements.

Using shared services, selected governmental functions can be concentrated into a single department or organizational unit and provided to other organizations based on agreed service levels. By unbundling and centralizing services, the basic premise is that services provided by one local department can be provided to others with relatively few efforts [1]. A shared services center might provide common services to local government organizations without affecting the autonomy of organizations and providing the flexibility to enhance and include additional functionality [5]. The management of relationship between shared service provider and requester seems crucial [4]. It is often unclear which shared service arrangements would be the best to leverage the full advantages of shared services. Moreover, often trade-offs are required, as one structure might facilitate efficiency over innovation, while another might stimulate high service levels over efficiency.

2. BACKGROUND

Shannon [8] defines simulation as the process of designing a model of a concrete system and conducting experiments with this model to understand the behavior of a concrete system and/or to evaluate various strategies for the operation of the system. Simulation can help to test and analyze different scenarios to understand their impact on a broader ‘system’ or provide ‘proof of concept’ evidence before moving forward with implementation plans [9]. In this way it enables managers to determine the likely consequences of investments, operational decisions and process changes before they are implemented. Simulation models describe behavior of the system under study and highlight factors that may result in failure and inefficiencies.

An important aspect of simulation is experimentation with alternative arrangements (e.g. [3]). The idea is to develop a model of the existing situation and based on a diagnosis develop one or more ‘to be’ models and find improved arrangements using what-if analyses. ‘What-if’ analyses using a simulation and animation model can be used to determine for example the consequence of changes in law and rules in the execution of public tasks.

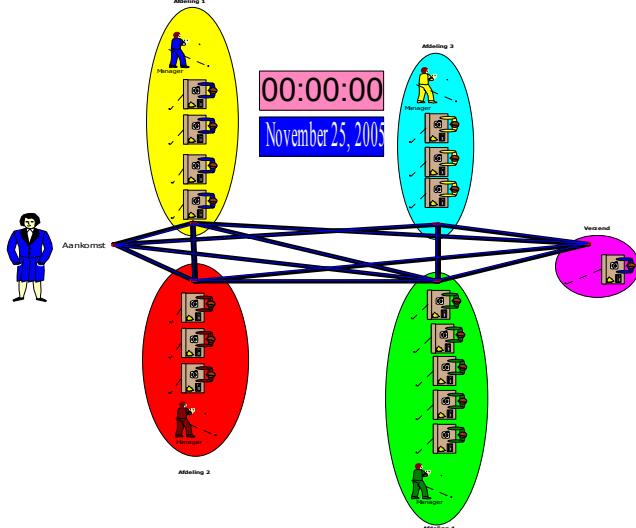


Figure 1. Screenshot of a simulation model

3. CASE STUDY

A case study of a municipality to be about to introduce shared services was investigated. The municipality was in the process of unbundling services from their local councils and concentrating it in one department. First, a model of the existing situation was constructed. In Figure 1 a screenshot of the animation of a simulation model of the situation without shared service centers is shown. The model shows that each department performs similar functions and makes the case for concentrating services in one department. The animation captures a number of views.

- *Application view:* the computer and applications on the desktop of employees are visualized. Also the interactions of users with the systems and interaction with other systems are visualized.
- *Business process view:* The time-dependent sequence of tasks performed by humans and computers is visualized through the flows of items. Work items are moving from the citizens to the departments and within departments.
- *Organizational view:* The departmental structure, citizens as customers, administrative workers and managers are visualized. The department are visualized using the colored circles.

Simulation and visualization was used to involve stakeholders, to enable discussions and to provide insight into the implications of possible shared service arrangements. A number of what-if analyses were conducted, including concentrating of all activities in one center and a mixed form where some services were centralized and other were kept decentralized. The alternative arrangements were compared with the existing situation based on a number of performance indicators.

The preliminary simulation results indicate that shared services arrangements result in a decrease of workers utilization and lower throughput time. The stakeholders valued the support positively, as the simulation and especially the visualization helped to improve the understanding of alternative arrangements and their impact. The, simulation helped to unravel the complexity of shared service arrangement.

4. CONCLUSIONS

In this ongoing research we used simulation to gain understanding of how shared services can be used to improve governmental operations. Simulating the organizational, business process, and application views was instrumental in the unraveling of potential shared services. Discrete-event simulation can help to analyze the current situation, evaluate various shared service arrangements and support determining feasible arrangements that may fulfill common objectives prior to implementation.

Currently we are in the process of collecting data and adding quantitative parts to the models based on an analysis of a real-life situation. In this way we should be able to determine the quantitative impact of shared service arrangements on aspects like efficiency, number of failures and adaptability to changes in law. This should help to address the ongoing discussion about centralization and decentralization. Moreover this research should result in a systematic approach for diagnosing potential shared services and facilitate organizational and process transformations needed for their implementation.

5. REFERENCES

- [1] Bergeron, B. *Essentials of Shared Services*, John Wiley & Sons, 2003.
- [2] Beynon-Davies, P. and Williams, M.D. Evaluating electronic local government in the UK, *Journal of Information Technology*, 18, 2 (2003) 137-149.
- [3] Janssen, M. and Cresswell, A. An Enterprise Application Integration Methodology for E-Government. *Journal of Enterprise Information Management*, 18, 5 (2005) 531-547.
- [4] Janssen, Marijn and Anton, Joha, Issues in Relationship Management for Obtaining the Benefits of a Shared Service Center. *Sixth International Conference on Electronic Commerce*, 2004, 219 - 228.
- [5] Janssen, M., and Wagenaar R.W. An Analysis of a Shared Services Center in E-government. *Proceedings of the Hawai'i International Conference on System Sciences*, January 5 – 8, 2004, Big Island, Hawaii, 2004.
- [6] Layne, K.J.L. and Lee, J. Developing fully functional E-government: A four stage model, *Government Information Quarterly*, 18, 2, (2001) 122-136.
- [7] McIvor, R., McHugh, M. and Cadden, C. (2002), “Internet technologies: supporting transparency in the public sector”, *International Journal of Public Sector Management*, 15, 3 (2002) 170-187.
- [8] Shannon, R.E. *Systems simulation: the art and science*, Prentice-Hall, 1975.
- [9] Sol, H.G. *Simulation in Information Systems Development*. University of Groningen, Groningen, The Netherlands, 1982.

Modeling and Forecasting of e-Vilnius Development

Arturas Kaklauskas

Institute of Internet and Intelligent Technologies

Vilnius Gediminas Technical University,

Sauletekio Ave. 11, LT-10223 Vilnius,

Lithuania

artka@st.vtu.lt

ABSTRACT

The investigation carried out by the author of this paper under "Intelligent Cities" and e-City projects helped to identify and describe major trends of e-cities development in industrialized countries as well as providing recommendations for e-Vilnius's development. Research included the following four stages: comparative description of the e-cities development in developed countries and in Vilnius, comparison and contrast of e-cities development in developed countries and Vilnius, development of some of the general recommendations as how to improve the efficiency levels for a e-Vilnius, submission of particular recommendations for the e-Vilnius.

Categories and Subject Descriptors

H.4.2. [Information Systems Applications]: Types of Systems – e-government applications.

General Terms

Simulation, Prediction, Sustainable e-Vilnius Development.

Keywords

Modeling, Forecasting, e-Vilnius Development.

1. Description of the Poster

Vilnius is the capital of Lithuania and one of the country's oldest cities. The honor of founding Vilnius is justly given to Gediminas (a Lithuanian Duke) in the year 1323. The capital is listed in the World Heritage Register of UNESCO. The population of Vilnius is 700,000. The investigation carried out by the author of this paper under "Intelligent Cities" (Framework 6 programme) and e-City (Phare PPF) projects helped to identify and describe major trends of e-cities development in industrialized countries as well as providing recommendations for e-Vilnius's development.

The research's aim was to produce an analytical model of the development of e-Vilnius by undertaking a complex analysis of micro, meso and macro environment factors affecting it and to present recommendations on increasing its competitive ability. The research was performed by studying the expertise of advanced industrial economies and by adapting it to Vilnius by taking into consideration its specific history, development level, needs and traditions. A simulation was undertaken to provide insight into creating an effective environment for the e-Vilnius development by choosing rational micro, meso and macro factors. The word 'model' implies 'a system of game rules', which the sustainable e-Vilnius development could use to its

best advantage. This research included the following four stages.

Stage I. Comparative description of the e-cities development in developed countries and in Vilnius: a system of criteria characterizing the efficiency of e-cities development was determined; based on a system of criteria, a description of the present state of e-cities of developed and transitional countries and Vilnius is given in conceptual and quantitative forms.

Stage II. A comparison and contrast of e-cities development in developed countries and Vilnius includes: identifying the global development trends (general regularities) of the e-cities; identifying e-cities differences between developed countries and Vilnius; determining pluses and minuses of these differences for Vilnius; determining the best practice for e-cities for Vilnius as based on the actual conditions.

Stage III. A development of some of the general recommendations as how to improve the efficiency levels for a e-Vilnius.

Stage IV. Submission of particular recommendations for the e-Vilnius was presented at this stage. Each of the general recommendations proposed in the third stage carry several particular alternatives.

In order to throw more light on the Model, a more detailed description of some above mentioned stages of analysis follows: identifying e-Vilnius developmental advantages and disadvantages, determining the best practice and some lessons to be learned from advanced e-cities.

Vilnius varies from the compared cities according to ICT infrastructure, institutional arrangements and government policy in ICT field, economic structure and functions, social, planning and legislative systems, institutions, traditions and cultures, economic performance, immigrant communities and other indicators.

After the analysis of Vilnius and compared cities it was established that Vilnius falls behind them according to a number of quantitative and qualitative indicators. Some of the indicators are provided below: ICT facilities are not that well developed; low level of ICT penetration in such sectors as education, health services, lower-tier municipal institutions; a fairly large informal economy; older generation is a much less active user of IS services and participant in IS in general; increased "brain drain" of ICT specialists; further deepening of IST as a separated, but not integrated into general socio-economic discourse field of activity and learning; lack of knowledge of teachers to use ICT in education; IST is treated as a separate subject only and not integrated into the learning process in

general (especially in secondary education); wide-spread believe in ability to solve all problems by administrative measures, disregard to market laws; insufficient pace of growth of ICT penetration in public sector and households; inefficient use of EU funds for IST projects; narrow scope of eGovernment services, a lack of detailed and sound public policy on the matter; an unstable and fragmented institutional framework for IS policy; a low rank of IS policy in comparison with traditional policies on the governmental agenda.

However, Vilnius has several advantages: dynamic development of IST and public interest in IST; new laws to leave behind administrative and intrusive regulation systems and to implement rules that support innovations, a variety of services and investments in telecommunications, data distribution, Internet services, eSignature services; liberalised telecommunication market; strong competition in mobile telecommunication market and in ICT services diminishes costs for end-users and offers new services; comparably cheap education of ICT specialists; diminishing costs for internet and hardware speed up the internet penetration in households, business and public sector; growing ICT market size and value reflects the orientation of industry towards more knowledge-based activities; expanding ICT-related sectors (telecommunications, IT industry and services) and increasing level of ICT usage in service sectors. Recognized need for expanding eGovernment services; private initiatives to support ICT infrastructure; participation in international ICT programmes; more efficient methods of e-learning; relatively low profit tax.

While implementing projects "Intelligent Cities" [4, 5], e-City [2] and other projects the experiences of different cities was analyzed by the author of this paper and it helped to determine the best practice for e-cities development for Vilnius. In order to throw more light on the best practices, a more detailed description of some examples.

Virtual communities portals have proliferated and these are broadly of 2 types: creating a virtual space for 'communities of interest'; on-line access to information and services for residents of specific geographical areas. Whilst the former can exist devoid of any shared physical space and bring together users who are separated by large distances and never likely to meet, the latter is built on the potential for 'real space and real-time' interaction of users. Its aim is to support users as they live in their cities and communities, enabling decision-making, providing another access route to services, and a forum for discussion [1]. New information technologies must not serve as another instrument of social or personal segregation. On the contrary, its possibilities must be exploited in all its potential to put into practice new ways of relationship, virtual communities, spontaneous groups of shared interests, forums of debate and participation. It is necessary to build a virtual city totally interrelated with the physical city and with a big density of contents and interactions. And must be constructed from the acknowledgement that this will be only possible spreading technological culture to the whole of society and creating spaces for the own society to build its own means of expression [7].

The project E-VOICE [3] intends to concentrate on e-democracy/e-government in order to try and renew the political information, communication and interaction processes between elected politicians, the administration and the citizens – including young people - on a local and/or regional level at various locations in the North Sea Region with the support of

the 'new' media (internet, e-mail, sms, i-mode, etc.) in combination with the 'old' media (television, radio, (mobile) telephone, newspapers, etc). Some possible examples are: the organisational development of digital office hours – citizens get the opportunity to pose questions to mayor, aldermen and/or council members by e-mail or by direct communication via the internet and web-tv; online townhall (e.g. experimental broadcasts of the yearly local-council budgetary meeting); digital debates and online panel discussions for citizens; electronic neighbourhood groups [3]. As a result of the delegation of various functions to the local districts and the resulting increased focus on a flatter, less bureaucratic structure in relation to decision-making processes it became necessary to develop new methods for use in local government [4].

In recent years, there have been growing demands for a more participative approach to societal decision making and a higher level of accountability on the part of politicians and decision makers. Concurrently, the development of the Internet has provided an infrastructure to achieve these ends through substantive e-democracy. e-Democracy systems have the potential to draw on developments in decision support systems (DSS), involving stakeholders and the public in societal decisions [6].

The following aspects were analyzed in this paper and the conclusions are as follows: e-Cities should be well informed of the micro, meso and macro environment levels in which they operate; e-Cities analyze the micro, meso and macro environment levels and distribute their resources to take advantage of the opportunities and to minimize threats to their activities; micro, meso and macro level factors can be optimized; model for e-Vilnius development was proposed; some global development trends (general regularities) of the e-cities development were identified; some general and particular recommendations how to improve the efficiency levels for a e-Vilnius were developed.

2. Selected References

- [1] Campbell, B., Slatcher, A., Birchall, P., Stephenson, K. (2004) Vision of the Regenerated, 'networked' Future City. Work Package 5. Intelligent Cities project. Framework 6 programme. Contract no.: 507860.
- [2] e-City. Contract No.: 2003/004.341.08.01.01.0001.
- [3] E-VOICE - the voice of the citizen in the information society - the challenge of future democracy in Europe. URL: <<http://www.vibamt.dk/Interreghome.nsf/0/bf86de719a39c87ec1256d050035f1db?OpenDocument>>, current as of March 7, 2005.
- [4] INTELCITY (Intelligent Cities), 5th Framework Roadmap Project. Framework 5.
- [5] Intelligent Cities project. Framework 6 programme. 2004-2005. Contract No.: 507860.
- [6] Niculae, C., French, S. Bringing understanding in societal decision making: explaining and communicating analyses? URL: <<http://www3.interscience.wiley.com/cgi-bin/abstract/108069484/ABSTRACT>>, current as of March 7, 2005.
- [7] Saragossa towards Knowledge Society. October 2003. URL: <<http://www.ayto-zaragoza.es/azar/ciudad/ciudad-conocimiento/ZTKS03.pdf>>, current as of March 7, 2005.

Modeling Online Participation in Local Governance

Andrea Kavanaugh

*Center for HCI
Virginia Tech
Blacksburg, VA
540-231-1806

kavan@vt.edu

Manuel Pérez-Quiñones*

540-231-2646

perez@vt.edu

Daniel Dunlap*

540-231-3121

dunlapd@vt.edu

Philip Isenhour*

540-231-3121

isenhour@vt.edu

ABSTRACT

This is a two-page highlights summary of our three-year digital government project funded by the National Science Foundation (September 2004-2007). We summarize here the highlights from the first year and a half of the project (September 2004-December 2005). In addition to ongoing analysis of requirements and current technology use, we have administered a random sample household survey (N=717), the first of two rounds, and we conducted a series of focus groups on political participation and information technology use. Our biggest challenges are in designing tools that aggregate dispersed discussion among citizens about local concerns. We have been particularly focused on blogging, and on ways to make blogs of interest to a user easier to find and to join and to share.

Categories and Subject Descriptors

K.4.2. [Computing Milieu]: Computing and Society – social.

General Terms

Measurement, Documentation, Design, Experimentation, Human Factors, Theory

Keywords

Digital government, community networks, political participation

1. INTRODUCTION

The current research is a comprehensive case study over a three-year period to model citizen and government use of technology, especially citizen-to-citizen deliberation and government integration of civic deliberation into decision-making, and the modification of innovative tools to facilitate and support these activities. We are triangulating quantitative and qualitative techniques, comprised of random sample household surveys, interviews (one-on-one and focus groups), session logs and a participatory Web forum. Our primary research objectives are:

- To model community use of network technology for group discussion and deliberation;
- To model local government use of network technology and the integration of citizen feedback into decision-making processes.
- To deploy and evaluate a suite of modified innovative tools for incorporating citizen deliberation into local government decision-making.

2. COMPLETED IN YEAR 1

The tasks of the first year included interviews with government representatives and citizens, survey questionnaire development and administration, one-day workshop with town and county government representatives, focus group interviews, current technology use and requirements analysis. Since the outset of the project in September 2004, we have been investigating current use of information tools by government staff and officials, and citizens and representatives of citizens groups. We have been conducting one-on-one and small group interviews with citizens groups (e.g., neighborhood association, political party affiliated group, local issue advocacy group) in order to learn about their interest in new capabilities or features for information technology that would help support or facilitate deliberation among group members. We have been conducting one-on-one and small group interviews with local government representatives (Town of Blacksburg and Montgomery County) regarding current technology use, problems and frustrations, and software capabilities and features that might help government become more aware of citizen discussion that is occurring online.

We have developed a telephone survey instrument to assess local political participation that we are administering in two rounds one year apart (April 2005 and April 2006) to the same households (total usable surveys for both rounds to be 500 respondents). The survey is designed to provide us with information about the general population as the context within which local deliberation is taking place. A subset of the population is engaged in deliberative activity at the local level. Where is deliberative activity occurring? To what extent is deliberation occurring within local formal and informal groups and voluntary associations versus ad hoc deliberation around issues of interest involving a dispersed general public? Our survey constructs include: Political Interests and Activities, Political Attitudes, Political Efficacy, Political Knowledge, Interpersonal Discussion Networks, Community Involvement and Attachment, Group Affiliations, Internet use for political participation and demographics [1, 4, 5, 6, 7].

In late summer 2005, we conducted a series of focus group interviews (the first of two rounds, one year apart) with a subset of survey respondents based on a diversity of respondent characteristics (level of political participation, political efficacy, Internet use, etc.). Among the tools we have been considering that might be helpful to citizens for deliberative activity are wikis and web logs (blogs). Commentary and deliberation on local issues is an interesting facet of the use of these tools. This is a fundamentally different model for civic deliberation than centralized, government sponsored efforts. Our ongoing work involves analysis of different models of deliberative activity. We have also been examining the value of RSS feeds linked to wikis and blogs. In addition to the many news websites such as Yahoo News and the BBC that have RSS feeds, recently some blogs and community sites have also been offering their content in syndicated format.

3. ONGOING IN YEAR 2

During the current second year, based on data collected during year one and ongoing interactions with community and government participants, we are undertaking the design, implementation, and deployment an initial set of BRIDGE-based prototype tools for online deliberation. During this year, the second government workshop and second round of citizen focus groups will include an overview and group discussion of the current state of these prototypes. The one-day government workshop in the second year will recruit at least some of the same ten representatives as the first year government workshop, plus at least five community leaders (including representatives from organizations serving lower socioeconomic groups, such as New River Community Action, and minority groups, such as the Hispanic and Black Caucuses of Virginia Tech). Community groups and government representatives will continue to use and evaluate the prototype tools during the latter half of the second year. We will conduct a second round of the telephone survey to provide longitudinal data and confirmatory analysis for our exploratory models of citizen-to-citizen deliberation online.

4. PLANNED FOR YEAR 3

In the third year we will finalize the data analysis from years 1 and 2 for the survey and citizen focus groups. We will continue making refinements to the BRIDGE-based prototype tools based on initial deployments in year 2, and will seek to expose more of the general public to these tools. We will conduct a participatory Web forum open to focus group and workshop participants, as well as the general public to evaluate these results and to comment on all strategies for increasing participation. Each year we will disseminate our findings through standard conferences and publications, as well as a project web site.

Our long term goals include better support for local government involvement, both as information provider and consumer of distributed intelligence produced in citizens' media-based deliberation. We seek to re-focus the digital government discussion around elements that make for an effective democracy rather than for effective government and in the participatory design and testing of innovative tools for lay citizens. The broader

impacts are both conceptual and practical as well as policy related. Our model of democratic deliberation that includes concrete tools could substantially modify future efforts to deploy this technology effectively by local government. This model could be particularly helpful in small towns and rural areas where technical expertise is scarce and infrastructure is sub-optimal.

Please see our project website for more detailed information:

<http://java.cs.vt.edu/public/projects/digitalgov>.

5. ACKNOWLEDGMENTS

We would like to express our deepest thanks to John M. Carroll, Mary Beth Rosson, and Joseph Schmitz for providing guidance and their expertise to the project. We would also like to thank Jaideep Godara, Matthew Cooper, Anshul Midha, Will Randolph, Salahuddin Hussein, Andrew Mike, B. Joon Kim and Alain Fabian for their assistance with this research.

6. REFERENCES

- [1] Carroll, J.M. and Reese, D. 2003. Community collective efficacy: Structure and consequences of perceived capacities in the Blacksburg Electronic Village. *Hawaii International Conference on System Sciences, HICSS-36* (January 6-9, Kona).
- [2] Coleman, S. and Gotz, J. 2002. Bowling Together: Online public engagement in policy deliberation; <http://bowlingtogether.net>
- [3] Fishkin, J.S. 1991. *Democracy and deliberation*. New Haven, CT: Yale University Press.
- [4] Kavanaugh, A., Carroll, J.M., Rosson, M. B., Reese, D. D. and Zin, T.T. 2005. Participating in Civil Society: The case of networked communities. *Interacting with Computers* 17, Special Issue on Designing for Civil Society, pp. 9-33
- [5] Kavanaugh, A., Isenhour, P., Cooper, M., Carroll, J.M., Rosson, M.B., and Schmitz, J. 2005. Information technology in support of public deliberation, pp. 19-40. In P. van den Besselaar, G. de Michelis, J. Preece and C. Simone (Eds.). *Communities and Technologies 2005*, The Netherlands: Kluwer Academic Publishers.
- [6] Kavanaugh, A., Isenhour, P., Godara, J., and Randolph, W. 2005. Detecting and Facilitating Deliberation at the Local Level. Forthcoming. In T. Davies and B. Noveck (Eds.), *Online Deliberation: Design, Research and Practice*. Chicago, IL: University of Chicago Press.
- [7] Kavanaugh, A., Rosson, M.B., Schmitz, J., and Kim, J. Forthcoming. Local Groups Online. In J.M. Carroll and M.B. Rosson (Eds.) special issue of the *Journal of Computer Supported Cooperative Work*.

Target Vehicle Identification for Border Safety with Modified Mutual Information

Siddharth Kaza, Yuan Wang, and Hsinchun Chen

Department of Management Information Systems

University of Arizona

1130 E. Helen St., Tucson, AZ 85721

1-520-621-2165

sidd@u.arizona.edu, ywang@email.arizona.edu, hchen@eller.arizona.edu

ABSTRACT

In recent years border security has been identified as a critical part of homeland security. The Department of Homeland Security monitors vehicles entering and leaving the country at land borders. Some vehicles are targeted to search for drugs and other contraband. Customs and Border Protection agents believe that vehicles involved in illegal activity operate in groups. If the criminal links of one vehicle are known then their border crossing patterns can be used to identify other partner vehicles. We perform this association analysis by using mutual information (MI) to identify pairs of vehicles that are potentially involved in criminal activity. Domain experts also suggest that criminal vehicles may cross at certain times of the day to evade inspection. We propose to modify the MI formulation to include this heuristic by using cross-jurisdictional criminal data from border-area jurisdictions.

Categories and Subject Descriptors

H.2.8 [Database Management]: Database Applications – *data mining*.

General Terms

Algorithms, Security

Keywords

Mutual information, Homeland security, Intelligence and security informatics

This research was supported in part by the NSF Digital Government (DG) program: "COPLINK Center: Information and Knowledge Management for Law Enforcement" #9983304, NSF Knowledge Discovery and Dissemination (KDD) program: "COPLINK Border Safe Research and Testbed" #9983304, NSF Information Technology Research (ITR) program: "COPLINK Center for Intelligence and Security Informatics Research - A Crime Data Mining Approach to Developing Border Safe Research" #0326348, and Department of Homeland Security (DHS) through the "BorderSafe" initiative #2030002.

1. INTRODUCTION

In recent years border security has been identified as a critical part of homeland security. The national strategy for homeland security [1] calls for the creation of "smart borders" that provide greater security through better intelligence. In addition, the report also emphasizes that information sharing systems are the foundations to improve the nation's infrastructure. The Department of Homeland Security (DHS) monitors vehicles entering and leaving the country, recording their license plates with a date and time of entry using license plate readers. These thorough checks are done for vehicles on watch lists (target vehicles) and on random vehicles as well. This process is time consuming and if the waiting times become too long, the flow of people, vehicles, and commerce is impaired. So, CBP agents are under pressure to balance security needs with efficiency. One of the aims of this study is to help CBP agents identify better quality target vehicles.

CBP agents believe that vehicles involved in illegal activity (especially smuggling) operate in groups. If the criminal links of one vehicle in a group are known, then the group's crossing patterns and frequency can be used to identify other partner vehicles. We perform this association analysis by using mutual information to identify pairs of vehicles crossing together and potentially involved in criminal activity. Our previous study [3] had found that the use of MI was a promising solution to this problem. In this study we modify the MI measure to incorporate domain heuristics. Domain experts (CBP agents, police detectives and analysts) suggest that groups of criminal vehicles may cross at certain times during the day to try and evade inspection. We use law enforcement information from border-area jurisdictions to identify times that criminal vehicles prefer and incorporate this knowledge in the MI formulation using conditional probability.

This study attempts to answer the following questions:

- Can law enforcement information from border-area jurisdictions be used to identify target vehicles at the border?
- How can we include domain heuristics to enhance the performance of mutual information?
- Which domain heuristics are important in identifying target vehicles at the border?

2. RESEARCH TESTBED AND DESIGN

The testbed for this study includes datasets obtained from the Tucson Police Department (TPD), Pima County Sheriff's

Department (PCSD), and Customs and Border Protection (CBP). These datasets are provided to us through the BorderSafe project funded by the Department of Homeland Security. The TPD and PCSD datasets include information on police incidents over 15 years (1990-2005). These incidents include individuals and vehicles recorded to be involved in illegal activity in a large part of Southern Arizona. A summary of these datasets is shown in Table 1.

Table 1. Key statistics of TPD and PCSD police data

	TPD	PCSD
Date Range	1990-2005	1990-2004
Recorded Incidents	3.3 million	2.18 million
Vehicles	800, 656	520, 539

CBP data includes information on vehicles crossing the border between Arizona and Mexico at six ports of entry. Details of this dataset are shown in Table 2.

Table 2. Key statistics of CBP border crossing data

Recorded crossings	10.7 million
Number of Vehicles	1.7 million

The MI [2] score between any two vehicles is defined as:

$$MIC(A, B) = \log_2 \frac{P(A, B)}{P(A)P(B)}$$

Here A is a vehicle with narcotics activity recorded in law-enforcement datasets, and B is a vehicle crossing within one hour of A . $P(A)$ and $P(B)$ are the probabilities of the vehicles A and B crossing the border. $P(A, B)$ is the probability of B crossing within one hour of A , this is calculated based on the number of times A and B are seen crossing together. Conditional probability is used to modify this formulation to include time of crossing and the percentage of criminal crossings in a time period. This modification gives more weight to crossings that take place during more criminal time periods (as identified by TPD/PCSD information).

3. EXPERIMENTAL RESULTS

MI was calculated for a total of 230,000 pairs of vehicles (the first vehicle from *Set A* and the second from *Set B*). To compare classical mutual information (MIC) and modified mutual information with time (MIT), we measured the number of criminal vehicles (as identified by TPD/PCSD) that each algorithm identified. These numbers are shown in Figure 1.

In Figure 1, on the X-axis are the top- n pairs ordered by MI scores. On the Y-axis is the number of criminal vehicles identified among the top- n pairs ordered by MI scores. So, from the chart, MIC identified 0 criminal vehicles, whereas MIT identified 2 criminal vehicles in the top-50 pairs. It can be seen that MIT consistently identifies more criminal vehicles than MIC in the top 2500 pairs. A pair-wise t-test for the difference in the numbers of criminal vehicles identified by the two methods was done. We found that MIT was significantly better than MIC at the 95% level in identifying criminal vehicles.

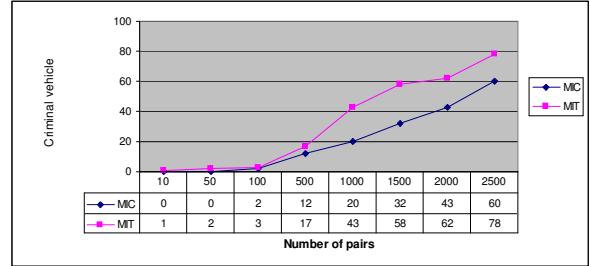


Figure 1. Number of criminal vehicles identified

4. CONCLUSIONS AND FUTURE DIRECTIONS

Exploring the criminal links of border crossing vehicles in local law enforcement databases can be used to enhance border security. In this study we used mutual information to identify pairs of border crossing vehicles that may be involved in criminal activity. We found that mutual information can be used to identify high quality potential target vehicles at the border. In addition, we concluded that the mutual information measure modified to include domain heuristics like time of crossing performs significantly better than classical mutual information in the identification of criminal vehicles. The method can be used to assist CBP agents to perform their functions both effectively and efficiently.

In the future, we plan to incorporate other domain heuristics in the mutual information formulation. We also plan to have domain experts from Customs and Border Protection validate our results and operationalize them.

5. ACKNOWLEDGEMENTS

We thank our BorderSafe project partners: Tucson Police Department, Pima County Sheriff's Department, Tucson Customs and Border Protection, ARJIS (Automated Regional Justice Information Systems), San Diego Super Computer Center (SDSC), SPAWAR, Department of Homeland Security, and Corporation for National Research Initiatives (CNRI). We also thank Homa Atabakhsh and Hemanth Gowda of the AI Lab at the University of Arizona, Tim Petersen and Chuck Violette of the Tucson Police Department, and Ron Friend of Tucson Customs and Border Protection for their contributions to this research.

6. REFERENCES

- [1] National Strategy for Homeland Security, Office of Homeland Security, 2002.
- [2] Fano, R.M. (1961) *Transmission of Information*. MIT Press, Cambridge, MA, 1961.
- [3] Kaza, S., Wang, T., Gowda, H. and Chen, H. Target Vehicle Identification for Border Safety using Mutual Information. In Proceedings of the 8th International IEEE Conference on Intelligent Transportation Systems (Vienna, Austria, 2005).

Friends, Foes, and Fringe: Norms and Structure in Political Discussion Networks

John W. Kelly

Columbia University

Communications Doctoral Program

New York NY 10027

+1 646 485 7364

kjw1@columbia.edu

Danyel Fisher

Microsoft Research

Community Technology Group

Redmond WA 98052

danyelf@microsoft.com

Marc Smith

Microsoft Research

Community Technology Group

Redmond WA 98052

masmith@microsoft.com

ABSTRACT

Online discussion groups have a network structure that emerges from the interactions of thousands of participants, writing in thousands of topical threads. This structure varies greatly according to the type of discussion group, such as technical, fan or support. Political groups have their own distinctive structure, organized around ideologically polarized clusters of participants. Whereas in some other groups, individuals who vehemently disagree with the mainstream might be ignored or ostracized, in political groups most participants preferentially interact with “opponents” and ignore “friends.” And yet, there is a type of opponent whose ideas are so far from the field of debate as to be ignored by most or nearly all other participants. This difference is starkly apparent in network diagrams of discussion groups. The core of highly participative discussants contains opponents from different ideological clusters, tightly bound in debate. But fringe contributors, sometimes called “trolls”, are relegated to peripheral positions by central actors’ lack of interest in responding to their provocations or views. Network visualizations of this phenomenon illustrate how macro-level structure arises and is maintained by micro-level discursive choices.

General Terms

Design, Theory.

Keywords

Public sphere, democracy, e-government, politics, online forum.

1. INTRODUCTION

The internet offers numerous modes of online discussion, with many different forms of control. Some empower one person to control agenda and content. Blogs are perhaps the most extreme version of this, in which one person contributes most of the content and can censor, delete or disallow feedback from others. Moderated discussion groups offer a less extreme version of such control, in which discussants are expected to carry on the majority of the discourse. Still other forums allow collaborative, group controls. Slashdot is a premiere example, in which users deploy randomly assigned rating points to grade particular comments up or down, making them more or less visible to subsequent readers [1]. If we envision a continuum of control, from the dictatorial blog on the one hand, through the constitutional monarchy of moderated discussion, to the kind of Athenian democracy (power being

randomly assigned to “citizens” for short durations) of Slashdot, the extreme anarchic pole is perhaps best represented by USENET [2].

Except in the case of a relatively few moderated discussions, USENET offers no overt forms of control to any participant. At most, one author can add disfavored others to their “killfile” and thus turn a deaf ear toward them. But they cannot diminish any other author’s access to the forum, and their only real power is to choose people to engage with, by deciding which posts to reply to. And yet, despite the “anarchy” of USENET, its various newsgroups feature stable, measurable structural characteristics. Somehow, order is maintained. And most interestingly, these regular structures vary greatly according to the social purpose of the newsgroup. For instance, a technical newsgroup, populated mainly with questions from the befuddled many and answers by the expert few, has a very different network profile than a support group, in which many regulars send welcoming messages to newcomers and there are broadly-distributed exchanges of advice and emotional solidarity[3].

Political newsgroups have their own distinctive network characteristics, and offer an interesting lesson in how regular structural features emerge from individual-level choices [4]. Despite persuasive speculation [5] and the tentative findings of some early internet research efforts [6], online political discussions need not necessarily become echo-chambers of the like-minded. The tendency to political homophily clearly exists in blogs [7] and seems to appear as well in more controlled environments featuring gatekeepers of one sort or another, but the kind of open, anarchic discussions found on USENET have quite the opposite tendency. We have previously found that debate, not agreement or reinforcement, is the dominant activity in political groups [8].

Consider the implications of a genre of discourse based around debate rather than information-sharing, emotional support, social coordination, or some other purpose. Clearly the latter sort of groups feature rather decisive forms of boundary-maintenance. In a technical newsgroup about Unix (for instance), someone offering a recipe for meatloaf would probably be ignored. Likewise someone posing as a Unix expert but offering fallacious advice would soon be identified as a charlatan [9], and likewise ignored. In a cancer support group, an author attacking the attitudes of other authors and offering detailed disputation of their posts would be

denounced and subsequently ignored by the community. In most newsgroups, antagonism and perceived wrongfulness are a ticket to rapid ostracism through the collective silence of the core author population. Only “newbies” speak to “trolls,” and only until admonished by more seasoned members of the newsgroup.

By contrast, it would at first blush seem like political newsgroups have no need of such boundary-maintenance. As we found previously, the great majority of authors (let us call them *fighters*) preferentially respond to messages from those on the other side; they respond to opponents more often than their allies. A second, smaller group of authors (we can call them *friendlies*) direct their attention to allies and refuse to engage opponents, despite the fact that they are routinely ignored by the former and harangued by the latter. Because their opponents do not reciprocate their discursive predilections by ignoring them, the *friendlies* are just as central to a political newsgroup’s core discussion network as the much more numerous *fighters*. In a political newsgroup, you cannot be left alone by the opposing cluster if you try. Indeed, it would seem that the only way to opt out of the fight is by opting out of the newsgroup altogether. But, interestingly and (to us) quite unexpectedly, there is a third type of author, even rarer, who tries hard not to be ignored, and nevertheless is. This type of author—the “fringe”—shows that boundary maintenance is at work in political newsgroups as well, and raises interesting questions for qualitative study.

We discovered this type of author serendipitously, while looking at ego network diagrams of core political newsgroup authors. In the following section we will take a look at some of these network diagrams, and see how they illustrate the link between authors’ micro-level choices about who to talk to, and macro-level structure of the discussion network. We also see boundary maintenance at work in an environment where most “enemies” are *good*, in the sense of being in demand, but not all enemies.

2. DISCUSSION NETWORKS

The current paper builds on the same data as our previous research [8], which contains a detailed account of the base data collection and analysis. In brief, core authors were identified for eight political newsgroups during November, 2003. Microsoft Research’s Netscan tool was used to capture a wide range of data on author behavior and thread structure, and was used to extract network data on core author behavior during the time frame. A *core author* is one who was among the twenty to forty most frequent (in terms of days active) contributors to the newsgroup during that month. A corpus of threaded political discussions was assembled containing hundreds of posts by all core authors. These were coded for evidence of political attitudes and for aspects of discursive behavior. Authors were clustered according to political attitudes, with only a small few found to be unclassifiable.

In the previous work, we showed that political newsgroups were found to have some distinctive features:

1. Almost all participants can be meaningfully assigned to distinct ideological or issue position clusters, depending on the particular newsgroup, for instance *left* and *right*, or *pro-choice* and *pro-life*.
2. Most newsgroups are bi-polar, in the sense of being organized around two dominant opposing clusters. In principle, some newsgroups could be multi-polar: one of the eight studied in [8] appeared to be centered around three dominant sides.
3. Replies to posts—and thus newsgroup interaction—is overwhelmingly across ideological or issue clusters, not within them.
4. Most authors choose to reply to messages by their opponents over their allies, and respond to far more messages on average from individual opponents than to individual allies. Further, [4] argues that political group members prefer to respond to people who are well-embedded in the conversation over new members.
5. Those rare authors who prefer to reply to allies are themselves, nevertheless, disproportionately

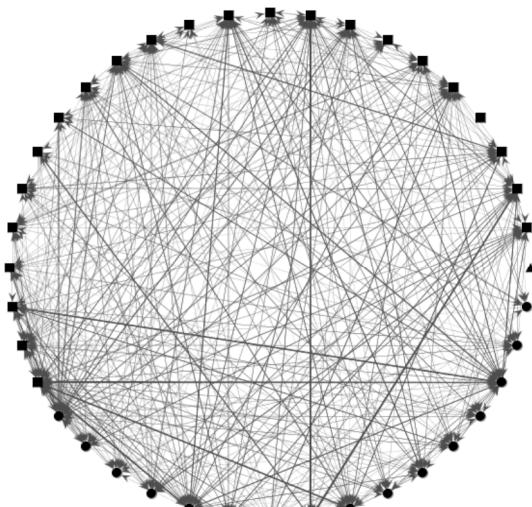


fig. 1 (link = 1 reply)

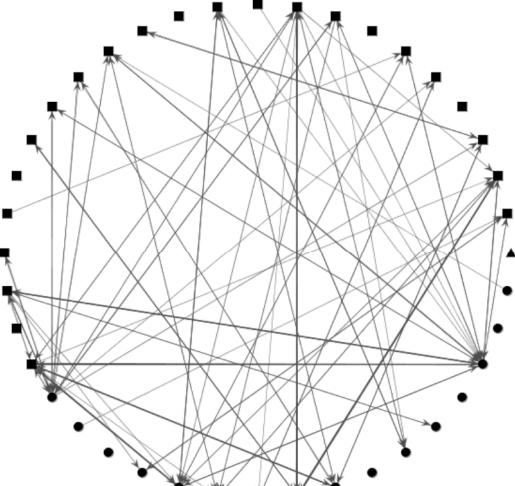


fig. 2 (link = 6 replies)

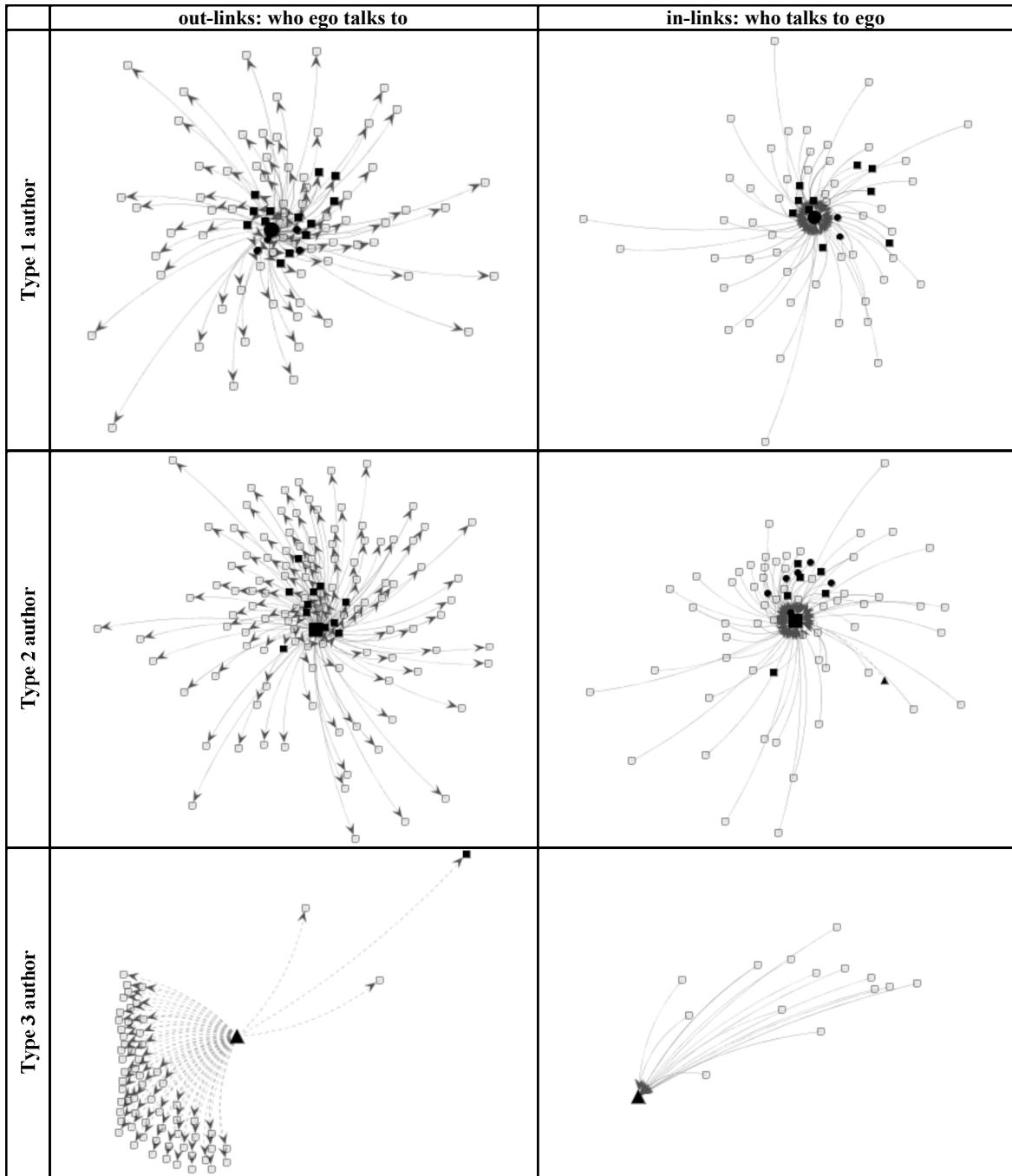


Figure 3: author choices and network response

responded to by opponents. Because of these authors, “in-links” (i.e. responses to an author by others) are very highly predictive of that author’s political position, much more so than their “out-links” (i.e. whom they choose to respond to).

6. There are tendencies toward balance in political newsgroups, in the following two patterns:

- a. Groups focused on a range of issues and featuring clusters best described as *ideological* (left/right, liberal/conservative, socialist/capitalist, etc) are generally balanced in both the populations of regular authors belonging to each cluster, and in the amount of message traffic generated by each cluster.
- b. Groups focused on a single contentious issue, like abortion or Middle East politics, are generally unbalanced in the population

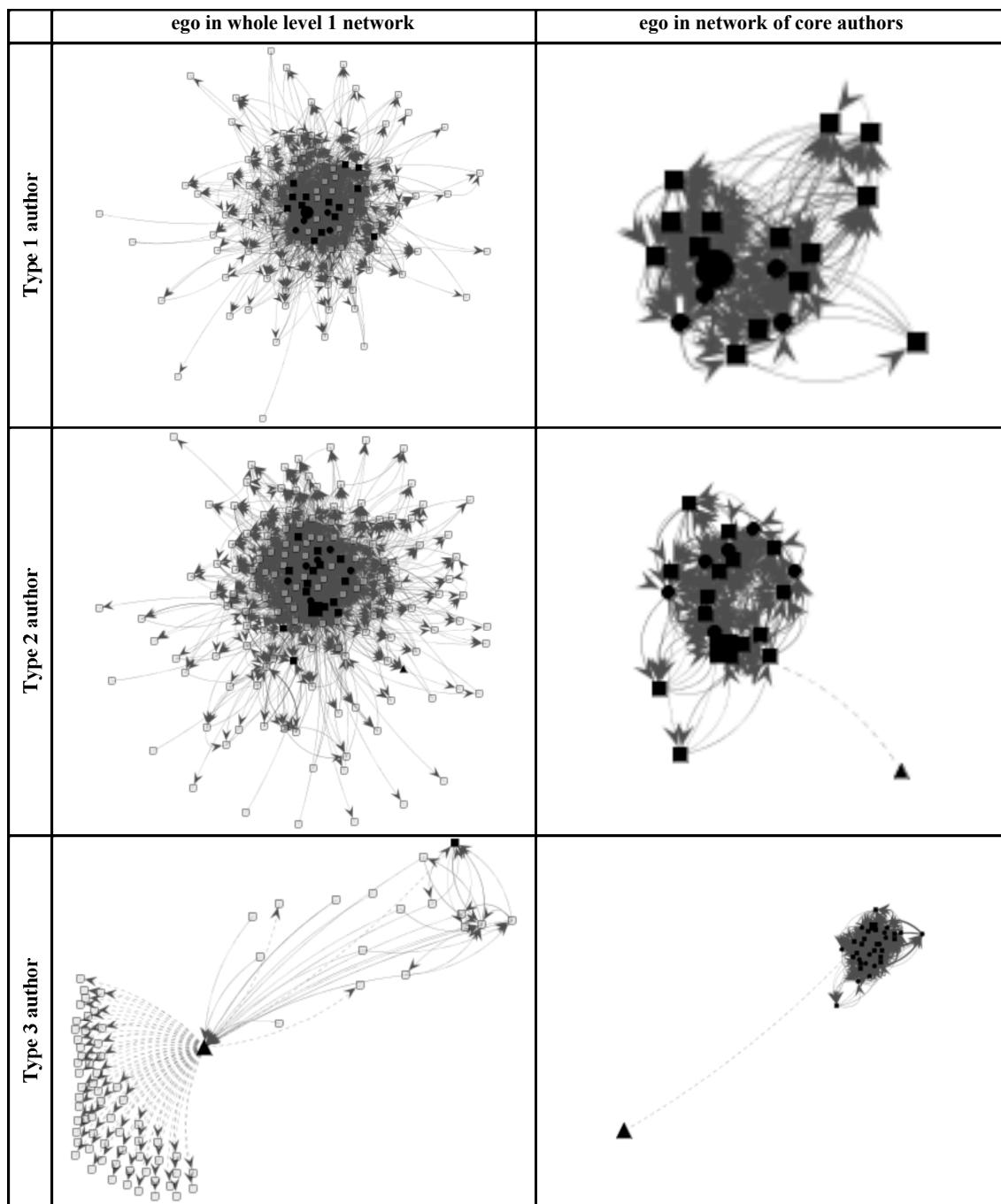


Figure 4: network position by author type

of authors belonging to each *issue-position* cluster, yet the minority authors post more messages on average and the message traffic generated by the clusters is thus significantly more balanced than the author populations.

As we will see in detail, these political and discursive tendencies yield a network structure in which an author population of discursive opponents, though politically clustered into two (or potentially more) distinct groups, are tightly bound in a central discussion core by dense bonds of

replies that tie opponents to one another more tightly than allies.

This does not mean that authors don't reply periodically to people who agree with them. If a network tie is considered to be a single reply, the core author population is so densely connected as to form almost a complete graph, i.e. a network in which all nodes are directly connected (Figure 1). To see the structure more clearly we must raise the number of replies that constitute a link, filtering out weaker bonds (Figure 2). Figures 1 and 2 show linked discussion cores from the newsgroup *alt.politics.bush*.

In those figures, nodes representing the core authors are laid out in a circle; authors who share a political position are placed near each other: liberals near other liberals (circles); conservatives near other conservatives (squares). Edges with arrows connect replies: an author “points to” another author by replying; more replies get a thicker edge.

Differences among types of political authors arise from their discursive behavior, and can be seen in a.) their choices about whom to reply to, b.) decisions by network members to reply to them, and c.) their position in the network structure arising from a and b (in combination with the same relationships among other actors in the network). In Figure 3 we can see micro-level features of author behavior, and the network’s response, for the three types of author. In those figures, too, persons holding opposing political positions are represented by squares and circles. Minor players—not in the core—are drawn as smaller gray shapes. A *fighter* (type 1 author) preferentially responds to opponents (out-links), and is likewise responded to mainly by opponents (in-links), and with only partial reciprocation from friends. The *friendly* (type 2 author) responds only to friends, most of whom do not reciprocate, and is responded to by a number of opponents anyway.

The *fringe* author exists at the edge of acceptable discourse within the group. Remember that the *fringe* author only shows up in the analysis because he is a regular contributor to the newsgroup, posting messages to it nearly every day. The *fringe* author’s views are extreme and do not fall into the newsgroup’s dominant ideological clusters (and so is coded as a triangle). This *fringe* author is a provocateur, posting a great number of initiating posts rather than replies. Many of the replies that he posts are “crossposts”: he replies to a message in a different group and adds this group to the conversation. (Crossposts are symbolized with dotted lines.) The author’s reply to a message by a core author (coded with a square) is ignored, and the only responses from the mainstream newsgroup population come from new and/or infrequent participants (“newbies,” coded light gray).

If we now turn from micro-level reply behavior to network structure, certain implications of that behavior are clear. The network diagrams of figure 2 use a so-called “physics model”: nodes repel from ones they are not linked to, and try to be a fixed distance from ones that they are linked to. Roughly, “close” suggests “likely to be connected”, while “far” suggests “less likely to be connected.” In these ego diagrams (ego is the biggest node), we can see that both *fighters* and *friendlies* are well-enmeshed in the discussion core. In fact, it is impossible to tell the difference between the two based on overall network position, because the replies to their messages are so dense. In contrast, the *fringe* author sticks out like a sore thumb. An author whose views are not seen as worthy of rebuttal or response by core authors is, figuratively speaking, and graphically visible, expelled from the network. Here we see boundary maintenance at work.

The group uses the one tool available to them, then, to maintain the boundary of “acceptable dialog”: they ignore this *fringe* author, giving him little satisfaction of triggering a broader discussion. Even in an arena dedicated to opposition—where every issue is contentious—the group comes to accord on what issues are not discussed, and leaves them behind.

3. CONCLUSION

Our example *fringe* author just one instance of the type. We have observed other *fringe* authors in different newsgroups, also far from the mainstream of debate. Their ego networks are similarly distinctive: they are isolated, garnering few responses from the active core of the newsgroup. Some of them attempt to reply more to core authors, some of them generate more or fewer seed posts, but all of them are relegated to the network periphery by the lack of demand for their ideas. What is very important to recognize, and very interesting, is that they are not marginalized because their ideas are uncomfortable, contentious, or, simply, disagreed with by others.

Keep in mind that most interaction, in fact the soul of interaction, in political newsgroups is strong, often vehement, disagreement between opponents. One finds Marxists sparring with Libertarians, liberal Democrats battling conservative Republicans, “pro-life” opponents of abortion calling “pro-choice” authors “murderers,” Israeli citizens arguing with Arab nationalists who think Israel should be pushed into the sea, etc. In USENET political newsgroups one finds people with strong and often irreconcilable views fighting each other in extended chains of argumentation. Sometimes it is emotional, with name calling of the worst sort. Sometimes it is highly rational, with detailed point-by-point rebuttals of quoted sentences and paragraphs. USENET authors seek out those with whom they disagree and expend enormous energy arguing with them. But the authors we here call *fringe* usually can’t get the time of day.

This behavior is noticeably different than that described by Baker [9]. Baker describes an amiable group, fans of a popular television show, who try to work over a period of several months to understand and change the behavior of an egregious “troll”. The group repeatedly engages the troll, responding to his posts and discussing his ideas, attempting to change his mind. Nowhere does he document a notion of ignoring the troll.

The reason for this requires further investigation no doubt, but is interesting to ponder. How might trolls and *fringe* authors be alike and how different? In some ways, the *fringe* authors behave like trolls, for instance posting incendiary messages and cross-posting their responses to messages into lots of other newsgroups. In other ways, including motivation, they may differ. Trolls often seem to be out to inflame other participants for the sake of being troublesome or disruptive, often appearing disingenuous or inauthentic to an experienced reader. By contrast, *fringe* authors in political groups usually seem quite sincere in their adherence to fanatical views. So, are *fringe* authors a type of troll? Or are both simply cases of bad citizens in the discursive community? Or are they very different types of actor altogether? In terms of behavior and motivation, and also network response, we should look more closely at *fringe* authors in relation to the more well-studied troll.

The *fringe* authors we have encountered are exactly the ones one would hope to find marginalized in a political discussion network. They are the sort who quote the “Protocols of the Elders of Zion” and offer genetic justifications for racial discrimination. Their views are not ignored because they are considered objectionable or extreme; indeed, extremity is often incorporated into the discussion. They are ignored because their ideas are not considered even mildly relevant to

any debate that anyone, on whichever side of whichever spectrum, wants to have. They are not even worthy of rebuttal. What people participating in political discourse care to discuss, as well as the particular attitudes they have about any given topic, are meaningfully related to the structure of concerns and attitudes in the larger political society to which they belong. In that larger society there are well-established political issues, frames and philosophies. To be involved in democratic life is to be engaged with these. People sometimes fear the internet as a political discussion medium. On the one hand it is accused of promoting smug, ideologically insular echo-chambers, and on the other it is said to hand the keys of the castle to Nazis, violent anarchists, and other assorted ideological bogeymen. But we should take heart from the findings of this study. In anarchic (in terms of rules of governance, not political philosophy) online political discourse networks, there is active boundary maintenance, informed by group norms held even among those who disagree strongly with one another about the topics under discussion. An author must be interesting to be engaged. The discourse network is shaped, and maintained, by *demand*, not *supply*. An implication of this is clear. What threatens democratic online political discourse and invites the worst sort of extremity is not the presence of radical voices, but the absence of reasoned ones.

4. REFERENCES

- [1] Lampe, C., and P. Resnick. 2004. *Slash(dot) and Burn: Distributed Moderation in a Large Online Conversation Space*. Proceedings of the 2004 Conference on Human Factors in Computing Systems (CHI 2004)
- [2] Pfaffenberger, B. 2003. A Standing Wave in the Web of Our Communications: Usenet and the Socio-Technical Construction of Cyberspace Values. *From Usenet to CoWebs: Interacting with Social Information Spaces*. London: Springer.
- [3] Turner, T., M. Smith, D. Fisher, and H. Welser. 2005. *Picturing Usenet: Mapping Computer-Mediated Collective Action*. Journal of Computer-Mediated Communication. 10(4), article 7.
<http://jcmc.indiana.edu/vol10/issue4/turner.html>
- [4] Fisher, D., M. Smith, H. Welser. 2006. *You Are Who You Talk To: Detecting Roles in Usenet Newsgroups*. Hawai'i International Conference on Systems Science, Kauai, Hawaii.
- [5] Sunstein, C. R. 2001. *Republic.com*. Princeton N.J., Princeton University Press.
- [6] Wilhelm, A. 1999. Virtual Sounding Boards: How deliberative is online political discussion. *Digital democracy : discourse and decision making in the Information Age*. B. N. Hague and B. Loader. London ; New York, Routledge: xvi, 277.
- [7] Adamic, L. and N. Glance. 2005. *The Political Blogosphere and the 2004 U.S. Election: Divided They Blog*. LinkKDD-2005, Chicago, IL, Aug. 21, 2005.
- [8] Kelly, J., D. Fisher, M. Smith. 2005. *Debate, Division, and Diversity: Political Discourse Networks in USENET Newsgroups*. under submission. Presented at the Stanford Online Deliberation Conference, 2005. available from <http://www.iserp.columbia.edu/coi/workingpapers.html>
- [9] Donath, Judith S. 1998. Identity and deception in the virtual community. In Smith, M and P. Kollock , Eds. *Communities in Cyberspace*. London: Routledge.
- [10] Baker, P. 2001. Moral Panic and Alternative Identity Construction in Usenet. *Journal of Computer-Mediated Communication*. Volume 7, Number 1.
<http://jcmc.indiana.edu/vol7/issue1/baker.html>

Electronic Government Capacity and Federal Program Performance: An Analysis of OMB's PART Scores and Executive Branch Management Scorecard

Hyun Joon Kim

Department of Political Science
Texas Tech University, Box 41015
Lubbock, TX 79409
1-806-742-4044

joon.kim@ttu.edu

Soonhee Kim

Department of Public Administration
Syracuse University, 306 Eggers Hall
Syracuse, NY 13244
1-315-443-1282

soonheekim@maxwell.syr.edu

ABSTRACT

This study analyzes the impact of e-government capacity and program management capacity on federal program performance by using two sets of performance data collected by Office of Management and Budget – the Performance Assessment Rating Tools and the Executive Branch Management Scorecard. The study results show that e-government capacity building creates a notable improvement in federal program performance. However, this effect was only observed when e-government capacity had reached a significant level of development. This finding implies that government organizations' investment in e-government should continue in order to develop e-government capacity in several critical areas: e-government planning, enterprise-wide architecture building, systems security assurance, implementation of projects on time and within budget, and provisioning services for diverse e-government service users. Consistent with the rationale behind the recent government reform efforts, this study also found that management capacities, including program design quality, program planning quality, and program management quality, are all significant antecedents of individual program performance.

General Terms

Management, Measurement, Performance

Keywords

e-government capacity, management capacity, organizational performance

1. INTRODUCTION

The purpose of this paper is to analyze the association between e-government capacity and individual program performance in the federal government. Drawing on the concept of management capacity, this study further explores the impact of program management capacity on federal program performance. For this empirical testing, we used two performance evaluation data sets collected by the Office of Management and Budget in 2004, which are the Performance Assessment Rating Tools (hereinafter the PART) and the Executive Branch Management Scorecard (hereinafter the Scorecard).

Using data from the PART and the Scorecard, this paper undertakes an empirical test of the proposition that e-government capacity improves federal program performance. This study also analyzes the impact of program management quality on individual program performance.

2. HYPOTHESES

This study is interested in the relationship between agency e-government capacity and individual program performance. We test whether the agency-level e-government capacity positively influences individual program performance in the agency. In addition to the performance of e-government, we also include three factors related to the program management process in order to identify their influences on program performance. These are program design quality, program planning quality, and program management quality.

H1: The level of e-government capacity is positively associated with program performance.

H2: The quality of program design is positively associated with program performance.

H3: The quality of strategic planning is positively associated with program performance.

H4: The quality of program management is positively associated with program performance.

3. DATA AND METHODS

Two different government data sets are used to test these hypotheses. We combined the PART data and the Scorecard data, both collected by OMB. The PART was developed to assess federal program performance by identifying the factors affecting program performance and defining the measures of program results.

The other set of data is the Scorecard. The Scorecard employs a grading system that shows whether a department or an agency runs its business successfully (green light), shows mixed results (yellow light), or performs unsatisfactorily (red light).

The most updated Scorecard data was reported in December, 2005. In order to match the time spans of the Scorecard data with the PART data collected by June 2004, we decided to use the Scorecard reported in June 2004. The PART data encompasses 42 departments and agencies, but the Scorecard

data includes the performance of 26 departments and agencies. Therefore we sorted out those agencies that did not appear in both data sets. Specifically, this study was based on the combined data set that has 571 programs' PART evaluation scores and the e-government Scorecard status scores collected from 24 departments and agencies.

The first three sections of PART scores are used to measure management process, including program design quality, program planning quality, and program management quality, and the last section of PART scores becomes the measurement of individual program performance. The PART data also includes the budget size of each program. Therefore, we also use the budget data for the control variable.

4. RESULTS

For estimation, we used Ordinary Least Squares. The Table 1 displays the estimation results. Both program planning quality and program management quality are found as highly significant antecedents of program results ($p < .001$), but the program design quality is significant at 90% confidence level. The estimation of e-government capacity's impact on program performance shows interesting results. When agency e-government capacity is graded as "success," i.e. green, the program performance is significantly better than other programs for which e-government capacity is considered moderate, i.e. yellow, if all other conditions are equal. However, the results do not support the hypothesis that programs whose e-government capacity reached the level of yellow perform better than programs at the red level of e-government, if every other condition is same. Therefore, the first hypothesis is partially supported by the estimation results and hypotheses two, three, and four are also supported.

Table 1. OLS Estimation Results

Variables	Est. Coeff.	Std. Error	p-value
Constant	-21.007**	4.666	0
E-government (green)	6.201*	3.349	0.065
E-government (red)	-0.649	1.754	0.712
Program Design	0.08*	0.045	0.076
Program Planning	0.655**	0.037	0
Program Management	0.161**	0.048	0.001
Budget Size	0	0	0.341
R square = 0.489; Adjusted R square=0.484; N=571 *P<0.10, **P<0.001			

We should note an important implication of the findings. The capacity of e-government may become an important antecedent of program performance only when the e-government capacity reaches a certain level. E-government capacity that remains below this critical level does not add a significant value to program performance. One may question what differences among three levels of e-government capacity resulted in the different impacts on performance. In order to answer that question, we need to compare the differences among the e-government Scorecard standards for each level. To be graded as green, an agency should show almost perfect compliance with

all the expectations. However, standards for the yellow grade are relatively loose compared with the green standards, and do not even include some of the items necessary to achieve a green standard.

For example, to receive a green grade, an agency must have a Modernization Blueprint for IT investments on important agency functions, but for yellow grade, an agency does not need to develop a blueprint. Having a well-defined IT investment plan may be one of the factors that make a difference in the overall e-government capacity of an agency. Another example of differences between green and yellow grades is that the strictness of timely management of IT projects within a given budget may cause a meaningful difference between green level e-government and yellow level e-government.

Furthermore, comprehensiveness in the areas of e-government projects also plays an important role. The agencies at green level should be conducting at least three of the four e-government initiatives. Since government programs may serve diverse service users – citizens, businesses, other governments, and internal agency users, balanced e-government development across multiple areas increases the likelihood of e-government's contribution to program performance enhancement. Security assurance of e-government systems can be another factor influencing the degree to which e-government affects program performance. Consistent, safe, and reliable operation of e-government systems may facilitate the maximum utilization of information resources necessary for the program implementation process. Accordingly, the findings imply that integrating various e-government capacity elements can make a meaningful difference in program performance.

5. CONCLUSION

The study results show that e-government capacity building provides a notable improvement in federal program performance. However, such an effect was only observed when e-government capacity has reached a significant development level. When e-government transformation is limited in its scope and premature in the degree of development, the expected effects of e-government capacity on performance enhancement is not likely to blossom to its potential. This finding implies that government organizations' investment in e-government should continue in order to develop e-government capacity in many critical areas: e-government planning, enterprise-wide architecture building, systems security assurance, implementation of projects on time and within budget, and provisioning services for diverse e-government service users.

Some limitations to this research should be noted. First, the programs included in the data set have various characteristics in terms of program type, duration, and recipients. Second, the effects of e-government capacity on individual program performance may be observable not only in the same time period but also across long periods of time. There may be a lag in the impact of e-government capacity on federal program performance. In spite of the limitations, this study can initiate and promote active discussion on future research ideas for e-government capacity and its impact on government performance for the e-government research community.

Research and Development for Innovative Government

– A National Agenda for Renewal

Trond Knudsen

R&D for iGovernment, Research Council of Norway

P.o.Box 2700 St. Hanshaugen

NO-0779 Oslo, Norway

Tel. +47 91703728

ABSTRACT

The Research Council of Norway (RCN) will present its initiatives to promote R&D for innovation in all levels of government. Highlights of relevant theoretical and empirical research on innovation in the public sector and innovation in services is briefly presented. An overview on the funding and management is given. Finally suggestions will be presented on R&D instruments for innovation in government, with ICT-based services in focus.

Categories and Subject Descriptors

[Digital Government]: Innovation research and development – *research instruments, ict based services*.

General Terms

Management, Measurement, Performance, Design, Economics.

Keywords

Digital government, eGovernment. Research instruments.

1. INTRODUCTION

The Research Council of Norway has a new obligation: to promote innovation in the public sector for the whole of Norway. This has been adopted as one of the strategic tasks of the Division of Innovation. An external committee has suggested the implementation of user driven R&D project support for need-motivated projects [1].

PUBLIN – Innovation in the Public Sector [2], a project funded by the EU's 5th Framework Programme, which recently presented its final results. These constitute an important basis for RCN's understanding of the nature of innovation in all levels of government. PUBLIN brings forward both empirical studies of innovation and innovation processes in government in several European countries and the theoretical background for understanding innovation as opposed to other changes and for describing different innovation types. The Research Council of Norway initiative focusing on R&D for innovative services, often technology enhanced, also is based on recent research in innovation services by ECON Analysis and Menon AS [3].

These two research based foundations bring forward suggestions for both coordination of selected R&D instruments and implementation of a new user driven instrument for meeting need-motivated R&D for innovation. Finally, the project "eGouvernet – The European eGovernment Network" financed by the EU 6th Framework Programme within Information Society Technologies as a coordination action starting January 2006 will contribute to a

common framework for eGovernment research in the future. Ministries and agencies financing eGovernment research from seven European nations together with EU's Institute for Prospective Technological Studies – IPTS, in Seville, Spain constitute the project's partners. An Interest Group of active non-partners is under constitution with actors and key players from public administrations, research and industry. This Group will contribute to the project's viability and the Research Council's overall strategy.

2. OVERALL OBJECTIVES

RCN's approach has the following overall objectives:

- To stimulate research based innovation in all levels of government
- To create a robust basis for technology enabled renewal
- To create a "Full Range" set of R&D Instruments for iGovernment
- To stimulate involvement of suppliers, NGO's, other relevant stakeholders and international R&D partners

3. BROAD IMPACT

Both industry and services are under rapid change, in part due to technology based disruptive changes. As shown by Econ Analyse and Menon, the innovation rates in services are high, and both production and distribution of services most often take new forms. This is believed to create both threats and possibilities for public sector:

- Preparing for the Demography Challenge in Health and Welfare
- Meeting Rising Expectations of Quality and Quantity of Public Service Production and Deliverance
- Developing the Supplier Sectors for Demanding Governmental Buyers at Home and Abroad
- Co-development of Technology, Organization and Co-operation for Work, Business and Services
- Contribution to Renewal of Governance and Democracy in the Global Service Society
- Broadband based services for regions and local governments

The need for mobilising research, suppliers, other stakeholders and governmental agencies and institution is accelerating as there still is a lack of resources put into this. The Norwegian government has also stating this need in its white Paper [4] on research:

- The Government: Research based innovation is needed for a major shift in efficiency and quality
- New instruments:
 - R&D for quality of service and integration
 - R&D for competence for innovation and renewal focusing
 - eGovernment
 - Efficiency for sustainable governance
 - Supporting the European arena for R&D for innovation
- Develop existing instruments:
 - Research for knowledge for policy development, focusing Europe in change, welfare, citizenship and democracy, migration and integration and more

In the PUBLIN Summary Report, it is stated:
Policy makers are experts in their own fields, and researchers will have to learn from them in order to understand the unwritten social, cultural and political context of policy development. Policy learning is therefore often the result of a fruitful interaction between policy makers and policy analysts. This perspective has importance for the development of learning and innovation networks and forums, where it can be useful to have members from both groups.

4. INNOVATION NEEDS

RCN is now discussing these possible topics for its support for R&D for public sector innovation in the years to come:

- Continuously mapping, analysis of instruments, programmes and strategies
- Identifying and coordination of complementary foci of different partners/stakeholders
- Establish framework for coordination between national programmes and EU FP instruments
- Some possible areas of interest:
 - Knowledge intensive services in public sector
 - Semantic interoperability
 - eHealth, telemedicine
 - Sustainable welfare services
 - Governance and democracy

5. RESEARCH INSTRUMENTS

Intensive work is going on to coordinate the use of several instruments for supporting R&D for innovation in the different divisions of RCN and with ministries and others. It is strongly believed that it is possible to be more successful in pursuing innovations by coordinating existing instruments and combining them with new instruments targeted at consortia research based in needs by agencies that will use the results. The following instruments are used today, and will profit from coordination in both thematic way and by alignment:

- Participation in FP5, FP6 eGov-initiatives
- Service-related R&D programmes: Local initiatives, 24/7 service, ALTINN, [Bønnøysund Registers Center](#), [Telemedicine](#)
- R&D programme: "Research for Innovation and Renewal in Public Sector" ([FIFOS](#))
- ICT-related R&D programmes: eHealth, tax form on the net, "Space", [eProcurement](#)
- With 'Innovation in Private Sector': Mobilizing Industry
- With *InnovationNorway*: Governmental R&D Partnerships and R&D in PPP
- RCN in-house collaboration: "Partnership R&D for Innovation in Public Sector" (VIOS) Establishing seamless R&D between generating knowledge for policy development and creating innovations in policy implementation

6. SOME CHALLENGES

It is a challenge to prove R&D for innovation within different levels and sectors of government, although some areas are more familiar with this paradigm. The health sector was an early adopter of new technology and new forms of organisation. However government is still, in general, mass producing services without profiting in full from technology enhanced tailoring combined with commercialisation and efficient mass production of knowledge intensive services. The reasons might be found among challenges like these:

- Establishing seamless R&D between generating knowledge for policy development and creating innovations in policy implementation
- Overcoming institutional barriers and competences for changes across sectors, systems and boundaries
- The process of building interdisciplinary innovation oriented R&D programmes with broad participation
- Initiate R&D for innovation and renewal of government synchronized with other relevant R&D and with the needs of national and international industry and

7. ACKNOWLEDGMENTS

Our thanks to researchers and analysts in the PuBLIN-project and its coordinator NIFU STEP, in econ Analyse and Menon AS and to colleagues focusing public sector and services research.

8. REFERENCES

- [1] Sandman, M. et al.: Virksomhetsforskret FoU for innovasjon i offentlig sektor - programforslag. *RCN Report* (Oslo, Jan. 2004) ISBN 82-12-01897-0, in Norwegian.
- [2] Koch, P.; Cunningham, Paul.; Schwabsky, N., Haukneset, J.: Summary and policy recommendations. *Publin Report No D24*. (Oslo 2006.) See <http://www.step.no/publin>
- [3] Econ Analyse, Menon AS: Innovation in Services *Publin Report No D24*. (Oslo 2006.) See <http://www.econ.no>

- [4] St.meld. nr. 20 (2004-2005) Vilje til forskning. (*In Norwegian*) See

<http://odin.dep.no/kd/norsk/dok/regpubl/stmeld/045001-040014/dok.bn.html>

A Distributed Information Management Framework (REGNET) for Environmental Laws and Regulations

Kincho H. Law¹

Professor of Civil and Environmental Engineering

Stanford University

Stanford, CA 94305-4020

Email: law@stanford.edu

ABSTRACT

The complexity, diversity, and volume of Federal and State regulations (as well as supplementary and supportive documents) are detrimental to businesses and hinder public understanding of government. The objective of REGNET project is to develop information infrastructure and tools for regulatory information management and to facilitate compliance assistance. As a pilot research application, the REGNET project focuses on environmental regulations. The experimental scope of this project focuses on Code of Federal Regulations (CFR) Title 40: Protection of the Environment and California Code of Regulations (CCR) Title 22: Social Security. Implementation examples include regulations and selected supplementary documents, covering hazardous waste, drinking water and the management of used oil. Furthermore, tools have been tested with additional environmental regulations from other States and regulations from other domain areas, such as CFR Title 21 on Food and Drugs, and different regulations related to accessibility from the US and UK.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval – *retrieval models*, I.2.1 [Artificial Intelligence]: Applications and Expert Systems – *law*, I.7.1 [Document and Text Processing]: Document and Text Editing – *Document management*.

1. INTRODUCTION

There has been a push by the executive office that government agencies put more emphasis on compliance assistance in lieu of enforcement to encourage companies to comply with regulations. It is well recognized that the complexity, diversity, and volume of Federal and State regulations are detrimental to businesses and also hinder public understanding of government. In addition to the regulations, supplementary and supportive documents (such as preambles, interpretation guides) are also an important part of regulatory information. The objective of REGNET project is to develop a formal information infrastructure for regulatory information management and to facilitate compliance assistance. As a pilot research application, the REGNET project focuses on environmental regulations.

The basic research tasks include: (1) textual parsing and storage, (2) semi-structured, indexed storage, (3) means to resolve semantic ambiguities, (4) cross-referencing appropriate for automated retrieval and analysis of relevant documents, and (5) on-line compliance checking of governmental regulations. The experimental scope of this project focuses on Code of Federal Regulations (CFR) Title 40: Protection of the Environment and California Code of Regulations (CCR) Title 22: Social Security. Implementation examples include regulations and selected supplementary documents, covering hazardous waste, drinking water and the management of used oil. Methodologies and tools have been tested with additional regulations, including environmental regulations from other States, CFR 21 on Food and Drugs and accessibility regulations from the US and UK.

2. CURRENT STATUS

2.1 Repositories and Access Tools

A textual parser has been designed and implemented to parse online HTML-based federal and state (California) environmental regulations into an XML structure. Specifically, the parser was successfully applied to parse the entire CFR 40 on Protection of Environment and CCR 22 on Social Security (and many other regulations.) In addition, the repository also includes selected supplementary documents dealing with used oil, which include the preamble to the regulation text found in 40 CFR 261 and 279, administrative decisions, guidance documents, federal cases, letters from the general counsel and letters of interpretation from the EPA. Online access tools have been built to allow searching and retrieval of regulations and documents. In short, we have developed a formal representation designed to handle regulatory provisions and documents as well as related supplemental information.

2.2 Ontology Development and References

One key issue dealing with a voluminous set of regulatory documents is to build appropriate ontology and concepts that can facilitate linking related regulations and documents stored in the repository. Tools have been developed to extract vocabulary and semi-automatically build structural thesauri (ontologies) for regulatory domains of interests. A parsing system has been developed using a context-free grammar and a semantic representation/interpretation system to automate the extraction

¹ Grant No.: EIA-0085998

CO-PI: Gio Wiederhold, Jim Leckie and Barton Thompson

Research Assistants: Jie Wang, Shawn Kerrigan, Gloria T. Lau, Bill Labiosa, Charles Heenan, Haoyi Wang, Xiaoshan Pan, Liang Zhou, Pooja Trivedi, Jun Peng

Institution: Stanford University

Website: <http://eil.stanford.edu/regnet>

of references and to tag regulation provisions with the list of references they contain. A research tool has also been developed to extract definitions on domain-specific terms and acronyms and attached them to the terms.

2.3 On-Line Compliance Checking

One objective of this research is to provide the means to interface the regulations with usages such that the regulations are not passive but active documents that can be dynamically linked to application programs for users to search and access regulations and to perform compliance checking. Depending on the nature of the regulations and compliance applications, two different approaches have been experimented.

The first is to express the rules formally in terms of First Order Predicate logic sentences and adopting (publicly available) theorem provers, a regulation assistance system has been designed and implemented to assist the compliance of provisions related to used oil management (40CFR260 and 40CFR279). The web-based compliance assistance system helps guide the user through the regulations, automatically insert links to any referenced regulation provisions, display terms and definitions and enable instant access to repository documents related to the provision. A demonstration system has been built for vehicle maintenance shops to check compliance with the used oil provisions.

Another approach is to directly link regulatory provisions with design applications, for instance, to ensure that a building design is in compliance with prescribed accessibility codes as specified in ADAAG (the Americans with Disabilities Act Accessibility Guide). A design-aid framework has been built to support compliance check of floor plans according to accessibility requirements and to conduct performance-based disabled access simulation developed using motion planning techniques.

2.4 Relatedness Analysis

As noted, legal regulations and information arise from diverse sources, each source has its own objectives, semantics, documentation format, and organization. It is not efficient, nor possible, to attempt to integrate multiple sources into a single consistent whole. One approach is to develop a tool to extract features and to compare similarities to determine the “relatedness” of the documents from different sources. Taking advantage of the ontology development and domain features, we combine knowledge composition and similarity analysis techniques and develop a relatedness analysis methodology. For example, the relatedness analysis tool is able to discover (almost exact) similarity between sections from Parts 141-143 of 40 CFR and Division 4 of 22 CCR on regulations related to drinking water. The tool has also been used to compare accessibility requirements among the US and the UK regulations. Furthermore, we have applied the prototype system to a E-rulemaking scenario where a set of proposed provisions (a 15-page document) with over 1400 public comments were processed to identify their relatedness. Preliminary results show the potential of the relatedness analysis tool to discover the public comments that are related to a specific drafted rule as well as those comments that are not relevant to the proposed rules.

3. RESEARCH PLAN

Preliminary studies have revealed the potential of the relatedness analysis approach to discover related regulations and documents. We believe the knowledge composition and similarity analysis approach is powerful and innovative, in that distinct knowledge sources or regulations do not have to be made completely consistent, only the terms that *articulate* their application connection are involved. The preliminary studies have, so far, been performed on a limited set of federal and state (drinking water) regulations and on a relative small set of comments available in another domain (accessibility) on an E-rulemaking scenario. Our research plan is to further develop and validate the “relatedness” analysis approach developed in this study and to demonstrate this approach for a broader set of federal and state regulations. Furthermore, new application of the relatedness analysis framework to develop “regulatory locator” for specific domains will be investigated.

At least two specific tasks have been identified:

- The first task is to further validate the relatedness analysis approach for larger sets of regulatory documents on a specific domain, for instance, to compare environmental regulations on hazardous waste management. Specifically, the federal regulations as described in Parts 260 to 265 of the 40CFR will be compared directly with Title 22 of CCR. These two sets of regulatory documents are significantly larger and more complicated than the regulations on drinking water used in the previous experimental study. Based on this experimental study, the relatedness analysis framework will then be refined and enhanced.
- The second task involves the application of the relatedness analysis framework to develop “regulatory locator” for specific domains. While regulations on a specific domain are mostly grouped under a specific title or part(s), related regulations also exist in other titles or parts. For example, “mercury”, a specific chemical, which appears in 40.CFR.141 (Part 141 of Title 40 of CFR), also appears in Title 21 of CFR on Drug and Food Administration. Similarly, the term “mercury”, which appears mostly in Title 22 of CCR, also appears in Titles 17 (Public Health), 8 (Industrial Relations), 3 (Food and Agriculture) and other parts of CCR. To fully locate “all” regulations for a specific hazardous substance is a very difficult task. Our plan is to apply our relatedness analysis framework and study the feasibility of extending the tool to facilitate the development of “regulatory locators”.

4. COLLABORATION AND OUTREACH

While this research focuses on the development of IT framework for regulatory information management and compliance assistance, the researchers have also been actively participating in the social and legal aspects of regulation compliance, enforcement and rule-making process. The researchers have participated in a number of workshops sponsored by Stanford’s Law School, Harvard’s Kennedy School of Government, EPA’s Office for Enforcement and Compliance Assurance related to IT and regulations. Active dialogues have been initiated with EPA, USGS, Pacific NW National Lab, Access Board and other agencies. We have also received software and hardware grants from Intel Corp., Autodesk Inc and Semio Inc..

Citizen Centric Analysis of Anti/Counter-Terrorism e-Government Services

H.R.Rao (PI)

Management Science and Systems
SUNY Buffalo
325C Jacobs Management Center
Buffalo, NY 14260
+1-716-645-3425

mgmtrao@buffalo.edu

JinKyu Lee

Management Science and Systems
SUNY Buffalo
248 Jacobs Management Center
Buffalo, NY 14260
+1-716-645-3425

jklee2@buffalo.edu

ABSTRACT

This paper presents an ongoing research project that seeks to leverage Information and Communication Technologies to better protect citizens from terrorism by enhancing citizen-government information flow. The paper summarizes results from two previous survey studies and presents a follow-up study planned in spring 2006.

Categories and Subject Descriptors

J.4 [Social and Behavioral Sciences]: Psychology

General Terms

Verification.

Keywords

Source choice, Channel choice, User acceptance, Anti-/Counter-Terrorism (ACT), e-Government, public safety

1. INTRODUCTION

The project seeks to leverage Information and Communication Technologies (ICT) to improve public safety and national security by facilitating information flow between citizens and Anti/Counter-Terrorism (ACT) authorities. This project was started because of the urgent and pressing needs to protect Americans and residents of USA in the face of increasing threats of terrorism.

2. CURRENT PROJECT ACTIVITIES

The project team has conducted two survey studies and plans to conduct one additional survey for the project. The first study was accepted in Decision Support Systems in its Special Issue on Cyberinfrastructure for Homeland Security (forthcoming 2006), and analysis of the second study results has completed. For the final survey, the project team is currently reviewing the details of the survey administrative process, including the randomness, geographic distribution, and target size of the survey sample.

The last survey involves two sample groups. The first group consists of ACT service subscribers of Terrorism Research Center (TRC), a partnering independent institute dedicated to research on

terrorism, information warfare and security, critical infrastructure protection, homeland security, etc., with a 5-million hits/month public website (www.terrorism.com). An online questionnaire survey will be administered to the TRC online service subscribers in March – April, 2006. The second group consists of general public recruited across the US through a market research firm. A paper-based questionnaire equivalent to the online questionnaire for TRC will be mailed to this group in May 2006.

3. PUBLISHED & ONGOING RESEARCH

3.1 Study #1

A developmental study [6] examines various factors that can affect citizens' acceptance of web-based Anti/Counter-Terrorism (ACT) services. Based on psychology-oriented economics theories of decision making under uncertainty [3, 8], the study synthesizes an analytical model of citizens' ACT service acceptance. The model includes 3 types of risks (i.e., privacy risk from the ACT authority, security risk from the Internet, terrorism risk) and counter beliefs that can reduce or cancel out the risks (i.e., domain competence, online competence, and good intention beliefs in the ACT authority and structural assurance belief in the cyberworld. Citizens' acceptance of a web-based ACT service is measured on two dimensions: intention to depend on the ACT information on a web site and intention to provide private information to the ACT authority. The data were collected through two surveys and analyzed using a Structural Equation Modeling (SEM) technique. The results show that perceived privacy risk from the ACT authority can significantly deter citizens' intention to provide ACT information such as public tips and leads for terrorism investigations. Also, the study found that citizens' belief that an ACT authority is competent in its domain (ACT operations) is the most critical factor for citizens' dependence on the ACT website for information. While perceived usefulness of a website exerts significant positive effects on both intentions, perceived risk of terrorism seemed to have no direct effect on the intentions. This study contributes to the research community in that:

The study fills the gap of knowledge by offering a citizen's perspective to e-government initiatives. In addition to service providers' perspectives that focus on how to implement information systems and share the data with other government agents, this study answers those questions as why some web-based e-Gov services are well accepted by citizens while some others are not.

The study synthesizes and empirically tests a model of citizens' web-based ACT service acceptance. This model can provide

researchers a strong foothold for a future e-Gov acceptance study, while the findings suggest some interesting future research topics.

The research paper from this study was accepted in Decision Support Systems in its Special Issue on Cyberinfrastructure for Homeland Security and will be publicly available in 2006.

3.2 Study #2

A follow-up study has been conducted in summer-fall, 2005, and a working paper is being developed. This study divides the issue of citizens' ACT e-Gov service acceptance into two distinctive preference problems: source choice and channel choice in order to clarify the reasons some citizens do not accept the FBI *Tips Online* service. Based on Theory of Planned Behavior [1], online trust theories [2, 7], and decision theories[4], citizens' intention to provide terrorism investigation tips to the FBI and intention to use the FBI *Tips Online* function are explained by perceived relative utility, subjective norm, behavioral control, and trusting beliefs. The results revealed that intention to use the FBI *Tips Online* function when citizens provide private information to FBI is determined by all three factors among which perceived utility is the strongest determinant. Furthermore, perceived privacy risk and belief in FBI's competence in online services have significant impact over and beyond perceived performance of the online function. In terms of the intention to provide private information to FBI, subjective norm and traditional concept of trust (i.e., good intention and competence in ACT domain) showed strong effects.

This study bears important implications for ACT authorities and the research community. First, the model departs from previous studies in that the source choice problem is separated from the channel choice problem. This approach can help researchers and practitioners pin-point the point of problem.

Second, the decomposed concept of trust [5, 6] supports that the traditionally assumed online trust effects do not exist in ACT service channel choice decision. According to the results, channel choice is indirectly influenced by citizens' belief in ACT authorities' competence in online services, while the widely referred concept of trust has direct and indirect impacts on source choice decision, with differing impact sizes. These findings will greatly improve ACT service providers' ability to improve citizens' acceptance of their services.

3.3 Study #3

The next study examines extended criteria of the two choice problems explored in Study #2: Source choice and Channel choice. In this study, perceived attributes of various ACT service providers (e.g., including FBI, TRC, State/Local Police) and service channels (e.g., TV, printed newspapers, Web sites, telephone, postal mail, email) are examined. The attributes of alternative ACT service providers include competence in the ACT service domain/Internet-based ICT operations/traditional service channel and institutional assurance of privacy protection. The attributes of alternative ACT service channels include channel reliability, availability, publication speed, and media richness.

These attributes are measured and analyzed following the Galileo measurement procedure. Factors that can influence the perceived attributes, such as messages from social networks & mass media, personal traits, demographic factors (e.g., cultural backgrounds, etc.), are also examined and analyzed using a SEM technique.

This study provides in-depth prescriptive knowledge that allows ACT e-Gov initiatives to leverage their organizational and

technological resources by re-positioning various ACT services. Individual ACT authorities can also benefit from this study by comparing desirable attributes for their services and their attributes perceived by citizens. Also, the antecedents of the perceived attributes will suggest effective ways to manipulate citizens' perceptions of those attributes and reveal the groups of citizens who have less benefited from or alienated by some ACT e-Gov services.

4. SUCCESSES, CHALLENGES, & FUTURE PLANS

This project has successfully conducted surveys which provided insightful information and knowledge (e.g., strong effects of privacy risk, e-service source and channel choice patterns), which have been made widely available through leading conferences and journals. However, the variety of ACT service types and the low likelihood of using an ACT service make it almost impossible to extend the study to actual behavior. This problem brings up another issue: measurement reliability and study scope. Exploratory e-Gov service acceptance studies often involve a number of latent variables, which quickly increase the volume of questionnaire. A researcher has to carefully decide where to draw the boundary in order to balance measurement reliability and implication of the study.

The project team is planning to extend the subject of this study to natural disaster management services, including FEMA, Local Emergency Responders (e.g., 911), and private mass media companies.

5. ACKNOWLEDGMENTS

This research has been supported by the National Science Foundation under grant 0548917. The usual disclaimers apply.

6. REFERENCES

- [1] Ajzen, I. The Theory of Planned Behavior. *Organizational Behavior and Human Decision Processes*, 50 (2). 179-211.
- [2] Gefen, D., Karahanna, E. and Straub, D. Trust and TAM in Online Shopping:An integrated Model. *MIS Quarterly*, 27 (1). 51-90.
- [3] Kahneman, D. and Tversky, A. Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47 (2). 263-291.
- [4] Kahneman, D. and Tversky, A. The Psychology of Preferences *Scientific American*, 1982, 160-173.
- [5] Lee, J., Kim, D.J. and Rao, H.R., An Examination of Trust Effects and Pre-existing Relational Risks in e-Government Services. in *Proceedings of the 11th Americas Conference on Information Systems*, Omaha, NE, 2005, 1949-1954.
- [6] Lee, J. and Rao, H.R. Perceived Risks, Counter-Beliefs, and Intentions to Use Anti-/Counter-Terrorism Websites: An Exploratory Study of Government-Citizens Online Interactions in a Turbulent Environment. *Decision Support Systems*. Forthcoming 2006
- [7] McKnight, D.H., Choudhury, V. and Kacmar, C. Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research*, 13 (3). 334-359.
- [8] Savage, L.J. *The Foundations of Statistics*. Wiley, New York, 1954.

Periodic Association Mining in a Geospatial Decision Support System*

Dan Li

Dept. of Computer Science
Northern Arizona University
Flagstaff, AZ, 86011
dan.li@nau.edu

Jitender S. Deogun

Dept. of Computer Science & Engineering
University of Nebraska - Lincoln
Lincoln, NE 68588
deogun@cse.unl.edu

ABSTRACT

This paper presents an approach for mining partial periodic association rules in temporal databases. This approach allows the discovery of periodic episodes such that the events in an episode are not limited to a fixed order. Moreover, this approach treats the antecedent and consequent of a rule separately and allows time lag between them. Thus, rules discovered are useful in many applications for prediction.

1. INTRODUCTION

Periodicity detection in time-related databases is a challenging data mining problem in many applications. Since a periodic pattern indicates something persistent and predictable, it is important to identify and characterize the periodicity. As part of an NSF supported Digital Government Research project, we are developing a Geospatial Decision Support System (GDSS), with an initial focus on drought risk management. One of our project objectives is to discover human-interpretable periodic patterns and rules associated with ocean parameters, atmospheric indices and climatic data.

Several algorithms have been developed for detecting periodicity in large datasets [1, 3, 4]. Most previous methods for periodicity detection are based on mining periodic symbolic patterns, where the occurrences of each item in a pattern have fixed order. Focusing on symbol sequences rather than time-related sequences limits the flexibility of algorithms. Many pruning strategies applied to symbol sequences for periodicity search cannot be applied to time-based sequences. In addition, most previous work on periodic association rule mining focuses on the discovery of periodic patterns and the antecedent and the consequent in a rule are merely determined by rule confidence once a periodic pattern is discov-

*This research was supported in part by NSF Digital Government Grant No. EIA-0091530, USDA RMA Grant NO. 02IE08310228, and NSF EPSCOR, Grant No. EPS-0346476.

ered. In this paper, antecedent and consequent episodes are treated separately for the purpose of prediction. The concept of time lag has been addressed in the previous work [2]. Here, this concept is extended to periodic pattern discovery.

2. GENERATING FREQUENT PARTIAL PERIODIC ASSOCIATION RULES

Apriori-like algorithms have been proposed recently to discover periodic patterns in sequential databases [1, 3, 4]. We develop another version of Apriori algorithm by integrating two properties discussed below:

1. Rather than only discovering symbolic periodic patterns in fixed order, our algorithm provides more flexibility by finding all periodic patterns with either fixed order or non-fixed order based on the concepts of parallel and serial episodes with user-specified window width. Since the problem of mining serial episodes is similar to the problem of mining symbolic patterns, we only address the problem of discovering periodic parallel episodes.
2. Unlike previous algorithms that divide time-related datasets into discrete time segments [1, 3, 4], our approach uses a sliding window to find all periodic patterns at any offset given a period p .

We define two data structures to accommodate sparse or dense input datasets. One is called *event-based linked list*; the other is called *window-based linked list*. For an event-based linked list, each event type in event sequences is associated with a two-dimensional linked list. The first dimension records window offset given a period p , and the second dimension records period number in which the event occurs. For a window-based linked list, each window offset is associated with a two-dimensional linked list. The first dimension represents period segments, and the second dimension records the events that occur in a particular period.

The event-based algorithm needs to store the actual occurrences of each frequent periodic episode, i.e., the linked list records all the periods where an episode occurs regarding a particular window offset. This could be a problem if events occur frequently. To overcome this, we apply the idea presented in [5] to the event-based algorithm, i.e., we only track the differences of occurrences between two episodes. This reduces the storage size dramatically for dense databases. An interested reader can refer to [5] for the details of implementation.

3. EXPERIMENTS AND ANALYSIS

Two data sets are generated to test the algorithms. The first data set is a synthetic data set which is constructed randomly. This data set includes events from 25 sequences occurring at 5000 time stamps. The second data set contains weather data which is collected at the automated weather station in Clay Center, NE, from 1950-1999, for drought assessment and drought risk management [2].

Figure 1 shows the algorithm computation time as the support threshold, min_sup , changes from 0.45 to 0.7 on the synthetic data. The system is implemented with three methods, event-based without diffset, event-based with diffset, and window-based methods. Generally, from this group of charts, we observe that: (1) The smaller the support threshold is, the more gap in system running time between the two event-based methods. This is because the smaller threshold results in more candidates, and in turn, more set intersection operations are needed to compute the support of a candidate episode. When event-based linked list is used to record the occurrences of each episode, the time used for set intersection increases sharply as the support threshold decreases. However, when we apply diffset to linked list, the space needed to save the linked list decreases on the synthetic data compared to original linked list. (2) The window-based algorithm runs faster than the event-based algorithm as the support threshold increases. Although the window-based algorithm needs more database scans compared with the event-based algorithm, the time used for the database scan is less than the time used for linked list storage and set intersection when the support becomes greater. (3) As the window width increases, so does the periodic episode generation time. This is reasonable because a wide window includes more events from multiple sequences, and in turn, generates more candidates.

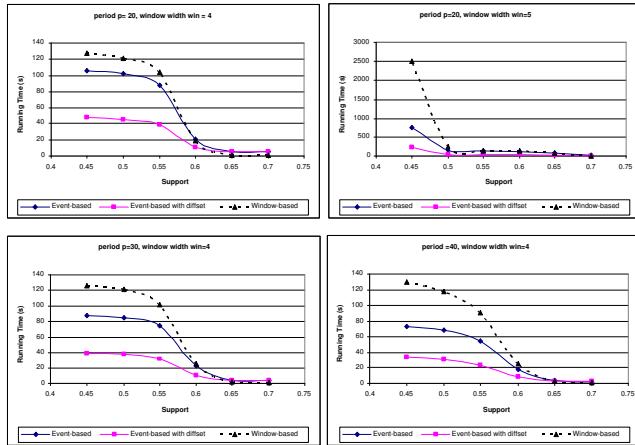


Figure 1: Running Time vs. Support.

Figure 2 shows the storage size on the weather data when two event-based methods are considered. Different from the results on the synthetic data, the experiments show that diffset requires more space to store the occurrences of episodes. This is because the events in weather data occur less frequently than the events in the synthetic data.

4. CONCLUSION

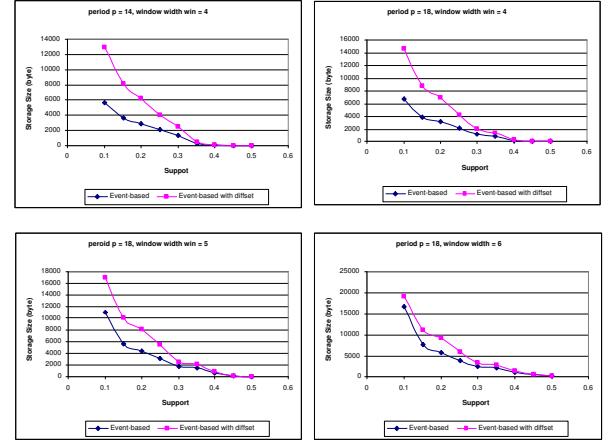


Figure 2: Storage Size vs. Support.

This paper presents the application of time-series data mining techniques in a Geospatial Decision Support System (GDSS) for drought risk management. We present efficient algorithms for mining partial periodic association rules in temporal databases. Two data structures, *event-based linked list* and *window-based linked list*, are examined. The strength of the event-based algorithm is that it needs only a single database pass. *Diffset* [5] is applied to the event-based method to reduce the storage space for dense datasets. The algorithm exhibits improved performance, especially when the number of frequent episodes is large. The window-based algorithm does not need extra space to save candidate episodes. Our experiments show that the event-based algorithm with *diffset* outperforms the window-based algorithm when we have a dense database in which events occur more frequently. The window-based algorithm demonstrates good performance when the value of user-specified support threshold is large. For a sparse database, the event-based algorithm without *diffset* provides the best results.

5. REFERENCES

- [1] J. Han, W. Gong, and Y. Yin. Mining segment-wise periodic patterns in time-related databases. In *Fourth International Conference on Knowledge Discovery and Data Mining*, pages 214–218, 1998.
- [2] S. Harms, J. Deogun, and T. Tadesse. Discovering sequential association rules with constraints and time lags in multiple sequences. In *Proceedings of the 2002 International Symposium on Methodologies for Intelligent Systems (ISMIS '02)*, pages 432–442, Lyon, France, June 2002.
- [3] Y. Li, P. Ning, X. S. Wang, and S. Jajodia. Discovering calendar-based temporal association rules. In *TIME*, pages 111–118, 2001.
- [4] B. Ozden, S. Ramaswamy, and A. Silberschatz. Cyclic association rules. In *ICDE*, pages 412–421, 1998.
- [5] M. J. Zaki and K. Gouda. Fast vertical mining using diffsets. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 326–335, 2003.

Digitalization of Coastal Management and Decision Making Supported by Multi-Dimensional Geospatial Information and Analysis

Ron Li, Keith Bedford, C.K. Shum,
Xutong Niu, Feng Zhou, Vasilis Velissariou
Department of CEEGS, The Ohio State University
470 Hitchcock Hall, 2070 Neil Avenue
Columbus, OH 43210
614-292-4303
li.282@osu.edu

J. Raul Ramirez
Center for Mapping
The Ohio State University
1216 Kinnear Road
Columbus, OH 43212
614-292-6557
raul@cfm.ohio-state.edu

Aidong Zhang
Department of Computer Science and Engineering
University at Buffalo, SUNY
201 Bell Hall
Buffalo, NY 14260-2000
716-645-3180 x 124
azhang@cse.buffalo.edu

ABSTRACT

This paper summarizes the results and outcomes of our NSF Digital Government project. During four years, this project concentrated on investigating, integrating and developing geospatial technologies to enhance the operational capabilities of federal, state, and local agencies responsible for coastal management and decision making. Successful collaborations have been established among research laboratories and government agencies in the areas of data collection and analysis, hydrodynamic modeling, development of web-based systems, and real-world applications of research results.

Categories and Subject Descriptors:

H.4 [Information Systems and Applications]: Spatial information technology

General Terms:

Management, Design, Verification

Key Words:

GIS, remote sensing, satellite images, decision making

This paper summarizes the results and outcomes of our project supported by the NSF Digital Government Program. The objective of this project is to investigate the current status, needs and potential of federal, state and local governmental operations related to geospatial information-supported coastal management and decision making. In the first three years, the project completed its tasks in Lake Erie. In the third and fourth years, we also used the developed strategy and technologies in the Tampa Bay project site where the coastal environment is different from that of the Great Lakes. Successful collaborations have been established among research laboratories and government agencies such as National Geodetic Survey (NGS)/NOAA, Office of Coastal Survey (OCS)/NOAA, Marine Geology Program/USGS, Tampa Bay Estuarine Program (TBEP), Southwest Florida Water Management District (SWFWMD), Ohio Department of Natural Resources (ODNR), and Lake County Planning Commission.

These collaborations were initiated in the areas of data collection and analysis, hydrodynamic modeling, development of web-based systems, and applications of research results [2]. They provided very useful advices on government coastal policies and plans, invaluable data sets, software for coastal modeling, and test site assistance.

A systematic study on generation of tide-coordinated shorelines [3] has been achieved. The proposed method for spatio-temporal modeling of tide-coordinated shoreline from historical shorelines made a significant step forward in terms of the theoretical representation and implementation. Two approaches to tide-coordinated shoreline modeling were developed.

The first uses instantaneous shorelines from satellite images and other sources to derive a tide-coordinated shoreline. A snake-based tide-coordinated shoreline model has been developed that can generate tide-coordinated shorelines by incorporating the internal forces coming from the shoreline geometry with external constraints such as tide gauge data, water-level change, coastal structures, and historic erosion rates.

The second estimates the tide-coordinated shoreline through a digital intersection between the digital coastal terrain model (CTM) and the digital water surface model (WSM). Digital shorelines have been generated in the Lake Erie area using an IKONOS-derived CTM along with three-, five-, ten-, and twenty-year hindcast water surfaces. This shoreline fits NOAA's aerial survey shoreline very well.

Water surface modeling and integration of multi-source coastal data is critical for coastal change detection [2, 5]. To develop and validate the hindcast capability of water surface modeling and then use it for future coastal change prediction, a set of twenty-year hindcast water surfaces (including mean lower low, mean, and mean high water levels) were produced from a hydrodynamic model for Lake Erie. The surfaces thus created were then compared with water-level observations from water gauge stations, satellite altimetry, 3-D shorelines derived from satellite images and aerial images, and other in situ data. Differences between the modeled water surface and the observations from various sources are around 20cm to 30cm. Considering the datum transformation errors in this area which are 18.2cm, the modeled water surfaces are accurate enough for the generation of tide-coordinated shorelines [3].

Based on the rational function model [1, 10], a multi-sensor integration between IKONOS and QuickBird stereo satellite

images has been studied [8]. Research results show that the 3-D geopositioning accuracy of the satellite stereo images is influenced by the intersection angle between two satellite imaging positions (same- or multi-satellite). A larger intersection angle gives a better level of accuracy in three-dimensional coordinates. A rational function model for aerial photographs has also been derived. With the addition of aerial photographs and a higher resolution in the multi-sensor integration analysis, the accuracy of 3-D positioning has improved significantly and can be used for better shoreline modeling.

We have developed a new method for the extraction of 3-D instantaneous blufflines from a combination of LIDAR and orthophotographs [4]. LIDAR can provide better vertical accuracy, though it has weak image texture information (especially for the horizontal position of blufflines). With the use of orthophotographs, the horizontal position of a bluffline can be easily identified. Thus, the combination of both LIDAR and orthophotographs produces 3-D blufflines with a better accuracy in both the horizontal and vertical directions.

A coastal decision-making system has been developed and enhanced by the addition of a PDA-based on-site mobile subsystem that utilizes portable GPS, wireless communications, and web-based GIS technologies [6, 7]. With this subsystem, a user is able to locate the exact position of coastal structures using GPS, and to access coastal data at a remote server from the field through wireless communications. Information about coastal structure permit applications can be inspected in the field by comparing the neighboring parcel data and the coastal environment using both remote GIS databases and the field data. At the same time, coastal engineers in the office can look at the same data just transferred from the field to make a joint decision.

An internet-based erosion awareness system [9] has been implemented for the use in Painesville, Ohio. Residents will be able to access this system to support activities such as buying or selling a house, or building erosion protection structures, while taking into consideration coastal change information. Local community leaders can use it for policy making and planning community activities.

In the fourth year, much of the research was focused on analysis of sea grass changes in Tampa Bay, FL. The major work includes:

- Using satellite, aerial and in situ observations, and hydrodynamic modeling techniques to monitor and model degradation of the sea grass population in Cockroach Bay, a small bay inside Tampa Bay. Historic sea grass coverage maps have been analyzed along with information on coastal environment change. Four different kinds of sea grass change patterns have been reviewed. Correlations have been studied between the spatial locations of these sea grass change patterns and changes in underwater slope.
- Determination and verification of the tide-coordinated shoreline model in a stretch of shoreline adjacent to Cockroach Bay section in Tampa Bay, FL. High-resolution CTMs have been generated from 2m water-penetrating LIDAR bathymetry (collected in 2004) and 2-5m QuickBird and IKONOS DEMs. Twenty-year water-gauge data have been acquired from NOAA and used in the Tampa Bay water surface model, in which tidal

influence has been studied. Digital shorelines have been generated through the intersection of twenty-year mean, mean lower low, and mean high water levels (computed from gauge stations) with the derived CTMs.

This research project has integrated the expertise and strengths of research in geographic information science and coastal engineering at The Ohio State University with database systems/computer science at State University of New York at Buffalo and government operations at federal, state and local government agencies. From the project results, we have demonstrated significant enhancement in the capabilities for handling of coastal spatio-temporal databases. A fundamental basis has been built for coastal geospatial information for inter-governmental agency operations. The system developed provides innovative tools for all levels of governmental agencies to improve efficiency and reduce operational costs. Finally, the project demonstrates the use of information technology and digital governmental operations for solving community problems in coastal property management.

ACKNOWLEDGMENTS

This project is supported by the Digital Government Program of the National Science Foundation.

REFERENCES

- [1] DI, K., MA, R., AND LI, R. 2003. Geometric Processing of IKONOS Geo Stereo Imagery for Coastal Mapping Applications. *Journal of Photogrammetric Engineering and Remote Sensing*, 69(8), pp. 873-879.
- [2] LI, R., BEDFORD, K.W., SHUM, C.K., RAMIREZ, J.R., ZHANG, A., AND DI, K. 2002. Digitalization of Coastal Management and Decision Making Supported by Multi-dimensional Geospatial Information and Analysis. In: *Proceedings of the NSF National Conference for Digital Government Research "dg.o 2002"*, Los Angeles, CA, May 20-22, 2002, pp. 53-59.
- [3] LI, R., MA, R., AND DI, K. 2002. Digital Tide-Coordinated Shoreline, *Journal of Marine Geodesy*, 25(1/2), pp. 27-36.
- [4] LIU, J., LI, R., AND SHIH, T. 2005. Estimation of Blufflines by Integrating Topographic LIDAR Data and Orthoimages (submitted to *Journal of Coastal Research*).
- [5] NIU, X., KUO, C.-Y., VELISSARIOU, V., LI, R., BEDFORD, K.W., AND SHUM, C. K. 2003. Multi-source Coastal Data Analysis. In: *Proceedings of the NSF National Conference on Digital Government Research*, Boston, MA, May 18-21, 2003, 227-230.
- [6] NIU, X., MA, R., ALI, T., AND LI, R. 2004. On-site Coastal Decision Making with Wireless Mobile GIS. In: *Proceedings XXth Congress of the International Society for Photogrammetry and Remote Sensing (ISPRS 2004)*, Istanbul, Turkey, July 12-23, 2004. (CD-ROM)
- [7] NIU, X., MA, R., ALI, T., AND LI, R. 2005. Integration of Mobile GIS and Wireless Technology for Coastal Management and Decision Making. *Journal of Photogrammetric Engineering and Remote Sensing*, 71(42), pp. 453-459.
- [8] NIU, X., ZHOU, F., DI, K., AND LI, R. 2005. 3D Geopositioning Accuracy Analysis based on Integration of QuickBird and IKONOS Imagery. In: *Proceedings of the ISPRS Workshop "High Resolution Earth Imaging for Geospatial Information"*, Hanover, Germany, May 17-20, 2005. (CD-ROM)
- [9] SRIVASTAVA, A., NIU, X., DI, K., AND LI, R. 2005. Shoreline Modeling and Erosion Prediction. In: *Proceedings of the ASPRS Annual Conference*, Baltimore, MD, March 7-11, 2005.
- [10] WANG, J., DI, K., AND LI, R. 2005. Evaluation and Improvement of Geopositioning Accuracy of IKONOS Stereo Imagery, *ASCE Journal of Surveying Engineering*, 131(2), pp. 35-42.

Locating Online Government Information: A Comparison of FirstGov, Google, and Yahoo

Lokman I. Meho and Kiduk Yang
School of Library and Information Science, Indiana University

1. Introduction

To facilitate both the understanding and the discovery of information, we need to utilize multiple sources of evidence, integrate a variety of methodologies, and combine human capabilities with those of the machine. The Web Information Discovery Integrated Tool (WIDIT) Laboratory at the School of Library and Information Science, Indiana University-Bloomington, houses several projects that employ this idea of multi-level fusion in the areas of information retrieval and knowledge discovery. This poster describes a comparative study of FirstGov, Google, and Yahoo by WIDIT's DGov research group, whose aim is to develop a more efficient and effective approach to organizing the U.S. Government websites that can facilitate the access and enhance the retrieval of government information.

2. GovSearch Project

As a first step in developing enhanced government information discovery system, we conducted a study to compare the retrieval results of FirstGov, Google, and Yahoo. The goal of the study was to identify the strengths and weaknesses of different search services so that we may be able to create a federated search system that maximize the strengths and minimize the weaknesses through data and method fusion. In keeping with the WIDIT DGov approach, which combines information retrieval (IR) and information organization (IO) methods to optimize the government information discovery process on the Web, we chose Google, which is a major IR service, Yahoo, which is a major IO service, and FirstGov, which was launched by Clinton Administration in 2000 to be the official gateway to government information.

To conduct our study, which is the first phase of the GovSearch project by WIDIT's DGov research group, we constructed a prototype federated search system called *GovSearch* that returns the search results from FirstGov, Google, and Yahoo in a side by side evaluation interface (Figure 1). Using the *GovSearch* system, we searched three search services with 73 Home Page finding queries (e.g. Food and Drug Administration), 73 Named Page finding queries (e.g. Free Application for Federal Student Aid), and 271 document finding queries (e.g. CIA World Factbook, the chapter/section on Turkey). The queries and relevance judgments for Home Page and Named Pages were selected from TREC-2004 Web track test collection (Craswell, 2004), and Document Page data was selected from a documentary source book (Meho, 2004).

Figure 1. *GovSearch* system: search and evaluation interfaces

The figure displays two side-by-side screenshots of the GovSearch Results interface. The left screenshot shows the search interface with a query field containing 'Food and Drug Administration'. The right screenshot shows the results page for this query, displaying three columns of search results from Google, Yahoo, and FirstGov respectively. Each result includes a snippet of text and a link to the full document.

Query #	Search Field	Match Type	Domains	Time	Google	Yahoo	FirstGov	Query Text
6	my	all	gov	13:50:58, 2004-11-05	1Y	1Y	20N	Philadelphia streets
7	my	all	gov	17:03:54, 2005-01-19	20N	20N	20N	Togo embassy
9	my	all	gov	14:37:22, 2005-01-20	20N	20N	20N	Baltimore
11	my	all	gov	14:38:48, 2005-01-21	1Y	1Y	20N	Karenna Ester

The *GovSearch* system queried the advanced search service of FirstGov (<http://www.firstgov.gov/fgsearch/>), the U.S. government search service of Google (<http://www.google.com/unclesam>), and the U.S. government category of Yahoo (http://dir.yahoo.com/Government/U_S_Government), using “all of the words” to find Home Pages and Named Pages, and “exact match” to find Document Pages. Top 20 results from each search services were manually evaluated as “YES”, “NO”, or “YLINK”, where “YLINK” indicated that the page contained a link to the relevant document. When the initial search did not return any relevant results, alternative search strategies were devised manually to override the default search.

The results of *GovSearch* study are as follows:

- Home Page finding task
 - 82 % of found items in FirstGov were ranked among the top 5, in comparison to approximately 95% in Google & Yahoo
 - 91% of found items in FirstGov were ranked among the top 10, in comparison to over 97% in Google & Yahoo.
 - FirstGov missed approximately one-fourth of the Home Pages whereas Google and Yahoo found all.
- Named Page finding task
 - 77% of found items in FirstGov were ranked among the top 5, in comparison to over 95% in Google & Yahoo.
 - 90% of found items in FirstGov were ranked among the top 10, in comparison to over 97% in Google & Yahoo.
 - FirstGov did not find 17.8% of the Named Pages whereas Google found all and Yahoo all but three documents
- Document Page finding task
 - Only Google is significantly successful in locating Document Pages, finding approximately two-thirds of all the documents.

3. Conclusion

The results of our study showed that FirstGov is not very effective in finding known items in government domain. It also showed that if an item is to be found using any of the three systems, it is generally found among the top 10 (with slight advantage of Google and Yahoo over FirstGov) and Only Google was able to locate a significant percentage of Document Pages. We believe that Google's success in retrieving Document Pages is partially due to the fact that it searches THOMAS, retrieving items from *The Congressional Record*. Google, however, still missed approximately one-third of the Document Pages, suggesting that searching and/or browsing in individual online government sources (e.g., GPO Access, National Archives and Records Administration, THOMAS, and Department of States' Web site) is still necessary to identify relevant information.

References

- Craswell,N. (2005). Overview of the TREC-2004 Web track. *Proceedings of the 13th Text Retrieval Conference (TREC 2004)*.
- Meho, Lokman I. (2004). *The Kurdish Question in U.S. Foreign Policy: A Documentary Sourcebook*. Westport, CT: Praeger Publishers. 720 p.

Interactive Design Best Practices for the Public Sector

Eric Miller

Squishymedia, Inc.

Portland, OR USA

+1 503.780.1847

eric@squishymedia.com

ABSTRACT

Generally speaking, best practices in the field of interactive design are often derived from private-sector practice or purely aesthetic considerations; the specific concerns of public-sector organizations are infrequently addressed. This poster series (produced by a web development practitioner) focuses on techniques for translating public sector organization objectives and requirements into specific 'real world' interface design practices. The conference presentation format is a visually oriented poster series derived from our past and current interactive design work for the public sector.

Categories and Subject Descriptors

H.5.2 [User Interfaces]: *Graphical user interfaces (GUI), Prototyping, Screen design, User-centered design*

General Terms

Design, Human Factors, Standardization

Keywords

Interactive design, Multimedia

1. INTRODUCTION

Interactive design projects (broadly defined as the design of communicative media in digital form, whether Web-based, kiosk, CD-ROM, mobile media, or physical installations) represent a complex field for interface design. Best practices are rapidly evolving within the field; existing design metaphors are often inadequate; and the immaturity of the technologies, platforms, and use conventions can hobble the effectiveness of a user interface.

In the case of public sector organizations, the difficulties can be compounded by additional design factors unique to the public sector and by a lack of clearly identified interactive design best practices specific to the public sector.

A growing consensus exists within the interactive design community emphasizing design as a process rather than simply a

product. Similar to the mechanical engineering design process defined by Mistree et al [1], interface design is viewed as a process through which the design team is translating a complex matrix of documentation and requirements into a communications strategy. In turn, the resulting communication strategy serves as a foundation for creating interface design parameters.

Our posters are designed to demonstrate a clear and coherent process-oriented methodology for interactive design based on the unique requirements of public sector development projects, and to assist dg.o 2006 attendees by providing insight and 'tips & tricks' that they might apply to their own projects in the future.

2. DESIGN PROCESS FACTORS

We have identified a few specific interactive design issues through our experience working with public sector clients.

- **Compliance with accessibility requirements** (Section 508/ADA) is mandatory for many government organizations, but the technology, practices, and relevant statutes are often inadequate or confusing
- Public media are expected to address the **broadest possible audience** since their audiences are often defined as 'the public' rather than a specific subset, which limits the use of design that leverages the visual communication conventions of the subgroup
- The complexity of the organizations, their mandates, and their messages can complicate the development of a comprehensive **information architecture** capable of coherently structuring the content
- Existing **technical infrastructure restrictions** from separate IS/IT organizations can limit the flexibility of the organization and/or implementation
- **Security requirements and public records laws** which may mandate specific techniques for handling content
- Staff **expertise and availability constraints** require that attention be paid to the longer-term sustainability of the final deliverables
- An **established 'house style'** design environment which may not be congruent with current design best practices
- A lack of focus on **branding and marketing** within the organization, whether intentional or unintentional, may detract from the overall communicative effectiveness of the design

3. DESIGN PRACTICE APPROACHES

Based on the previous list, the poster presentations contain written and visual examples of techniques to address each of these key areas of concern.

- **Compliance with accessibility requirements** through the use of XHTML, ADA-friendly technology selections, accessibility validators, and the use of specific coding techniques for assisting disabled users
- **Audience-specific design** by defining design requirements through user-centered design practices such as those articulated by Katz-Haas [2]. Defining general communication requirements and finding ways to communicate those elements without intruding on the effectiveness of the user-centered design.
- Developing the **information architecture** through collection and analysis of content, followed by the creation of a conceptual hierarchy and/or taxonomy. Adopting design approaches drawn from Garrett's Elements of User Experience [3]. Defining the process by which the information architecture will be translated into design parameters.
- Understanding **technical infrastructure restrictions** beforehand through communications with the technical staff and a structured assessment of the environment
- Compliance with **security requirements and public records laws** through a structured process of requirements identification, impact assessment, definition of compliance, and development process documentation, and translation into specific interface design parameters. Confirm that the use of licensed software (whether commercial or open-source) complies with established policy.
- Define implementation approaches and final deliverables that incorporate the **expertise and availability constraints** within the organization, such as compatibility with WYSIWYG production tools, minimizing use of compiled components such as Flash or Java, and/or implementing Wiki or CMS framework based solutions
- Leveraging the **established 'house style'** design by identifying which components are key to the organization and which elements might contribute most to the user-design centered methodology. Identify specific valuable components of the house style, then apply a 'Tufte Checklist' drawn from Tufte's

Envisioning Information [4] to determine how these elements can be adapted for reuse

- Identifying basic **branding and marketing** guidelines based on organizational communication objectives, then documenting the theory and practice behind the design's implementation of those guidelines for future use in the form of a style guide or branding manual

The examples used in the posters have been drawn from our current and prior public sector interactive design projects as well as other interactive design projects familiar to the digital government community.

4. VISUAL DISPLAY

The poster series produced for dg.o is itself designed to be an embodiment of the process outlined here. We have produced a poster presentation composed of a series of panels, each outlining a specific interface design challenge and providing visual examples of design solutions based on our experience and our prior digital government design projects.

In addition to discussing general design issues faced by public sector organizations, the material is designed to provide dg.o 2006 attendees with a concrete understanding of a structured design process relevant to their projects that require the services of visual designers and interactive developers.

Posters and related materials are available for download:
<http://www.squishymedia.com/docs/diggov2006/>

Our thanks to Portland State University for providing printing services for the poster presentation.

5. REFERENCES

- [1] Mistree, F., Smith, W.F., Bras, B., Allen, J.K., and Muster, D. "The Decision-Based Design: A Contemporary Paradigm for Ship Design," *Transactions, Society of Naval Architects and Marine Engineers*, Vol. 98, 565-597, 1990.
- [2] Katz-Haas, R. "Ten Guidelines for User Centered Web Design," In *Usability Interface*, Vol. 5, No. 1, 1998.
- [3] Garrett, J.J. The Elements of User Experience (.pdf), <http://www.jjg.net/ia/>, 2000.
- [4] Tufte, E. *Envisioning Information*, Graphics Press, Cheshire, CT, 1990.

Multi-Institution Testbed for Scalable Digital Archiving

Stephen P. Miller, Scripps Institution of Oceanography
Robert Detrick, Woods Hole Oceanographic Institution
John Helly, San Diego Supercomputer Center

The Scripps Institution of Oceanography (SIO) and the Woods Hole Oceanographic Institution (WHOI) have joined forces with the San Diego Supercomputer Center to build a testbed for multi-institutional archiving of shipboard and deep submergence vehicle data. Support has been provided by the Digital Archiving and Preservation program funded by NSF/CISE and the Library of Congress.

In addition to the more than 92,000 objects stored in the SIOExplorer Digital Library, the testbed provides access to data, photographs, video images and documents from WHOI ships, Alvin submersible and Jason ROV dives, and deep-towed vehicle surveys. An interactive digital library interface allows combinations of distributed collections to be browsed, metadata inspected, and objects displayed or selected for download. The digital library architecture, and the search and display tools of the SIOExplorer project, are being combined with WHOI tools, such as the Alvin Framegrabber and the Jason Virtual Control Van, that have been designed using WHOI's GeoBrowser to handle the vast volumes of digital video and camera data generated by Alvin, Jason and other deep submergence vehicles.

Notions of scalability will be tested, as data volumes range from 3 CDs per cruise to 200 DVDs per cruise. Much of the scalability of this proposal comes from an ability to attach digital library data and metadata acquisition processes to diverse sensor systems. We are able to run an entire digital library from a laptop computer as well as from supercomputer-center-size resources. It can be used, in the field, laboratory or classroom, covering data from acquisition-to-archive using a single coherent methodology. The design is an open architecture, supporting applications through well-defined external interfaces maintained as an open-source effort for community inclusion and enhancement.

Repository Replication Using SMTP and NNTP

Michael L. Nelson, Joan A. Smith, Martin Klein

Old Dominion University

Department of Computer Science

Norfolk VA 23529 USA

+1 757 683 6393

{mln,jsmith,mklein}@cs.odu.edu

ABSTRACT

We describe our progress on NSF ISS 0455997, "Shared Infrastructure Preservation Models". The focus of our efforts is to evaluate different preservation models based on Internet infrastructure that sites already have. Specifically, we investigate replicating the contents of a repository using the Simple Mail Transport Protocol ("email") and the Network News Transfer Protocol ("news").

Categories and Subject Descriptors

H.3.7 [Digital Libraries]: Collection, Dissemination

General Terms

Algorithms, Measurement, Performance, Design, Reliability, Experimentation.

Keywords

Digital Preservation, Internet Protocols, OAI-PMH, SMTP, NNTP.

1. INTRODUCTION

We are interested in digital preservation models that rely on shared, existing infrastructure. The premise is that if archiving can be accomplished within a widely deployed infrastructure whose operational burden is shared among many partners, the resulting system will have only an incremental cost and be tolerant of dynamic participation. We are investigating models that use network news transfer protocol (NNTP or "Usenet news") and simple mail transfer protocol (SMTP or "email"). We had originally hoped to include a multicast model, but upon further reflection we decided multicasting would require more infrastructure support than we were aiming for.

Through shared infrastructure, the individual cost for participation is kept low and aggregate system functionality is not greatly impacted by participants joining and leaving. The models primarily focus on creatively utilizing existing Internet protocols and systems in order to decrease the deployment and maintenance burden for participating organizations. Our models will not "automatically" preserve

data, but they should make it easy for those who wish to do so. We evaluate our models in the context of Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) repositories [1], although any repository infrastructure can be used.

2. EMPIRICAL EVALUATION

For our test repositories, we are using test websites that have the "mod_oai" Apache module that allows OAI-PMH access to web sites [2]. Unlike most OAI-PMH repositories, mod_oai allows for the entire resource to be disseminated via OAI-PMH, and not just metadata about the resource. Thus, mod_oai can be used to unambiguously serialize an entire web site in a series of OAI-PMH "ListRecords" responses. Our test web sites consist of 72 files (3 directories of 24) with file sizes ranging from 1.7KB to 1.5MB bytes.

An OAI-PMH harvester harvests from the test repository and then communicates the results to an instrumented SMTP or NNTP service.

2.1 NNTP SERVERS

We have created newsgroups that correspond to individual repositories. A harvester takes the OAI-PMH responses, and posts a base64-encoded version to the appropriate newsgroup (Figure 1). The messages traverse the existing NNTP network and can be retrieved at a later date by subscribers. The individual newsgroups are moderated to prevent anyone but the OAI-PMH harvester from posting.

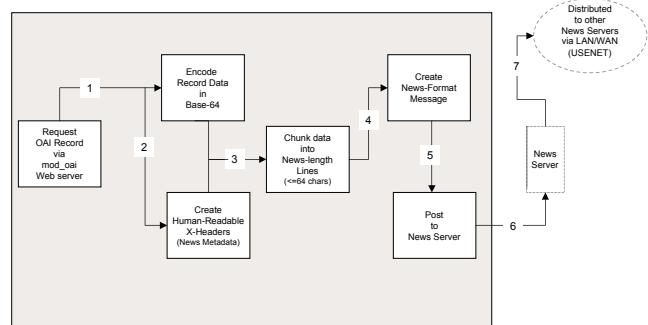


Figure 1. Posting OAI-PMH Responses as News

2.2 SMTP SERVERS

We have modified the Postfix mail daemon to intercept outgoing emails and attach OAI-PMH records to them (Figure 2). If a recipient can process the attachment, the OAI-PMH record is stripped off and saved for post-processing (Figure 3). Early results show a linear relationship between time required

to process attachments through our filter and the attachment size. Attaching

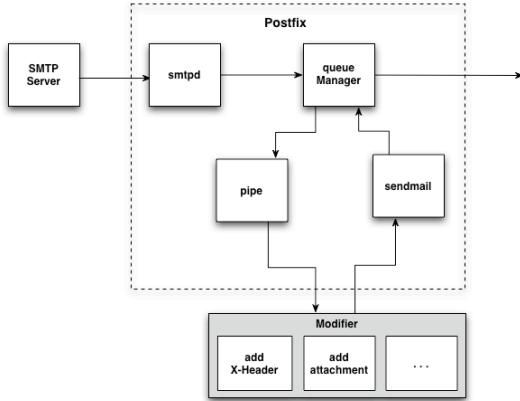


Figure 2. Modified Outbound Emails

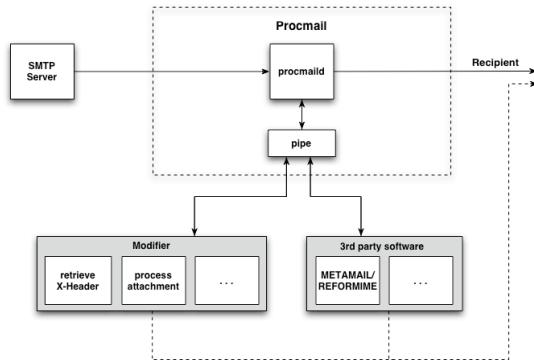


Figure 3. Modified Inbound Emails

3. SIMULATION

Based on the values obtained through instrumenting SMTP and NNTP servers, we are designing simulations to explore which policies would be best suited for repositories different of different sizes.

3.1 OAI-PMH Parameters

- R = # of records in the repository
 - R_s = mean size of records
 - R_u = # of records updated per time unit
 - R_a = # of records added per time unit
- Special cases:
- by reference: $R_s = 100$ bytes
 - "advertising": $R=1$, $R_s=100$ bytes (this is just a baseURL or Identify response)

3.2 NNTP Parameters

- t_b = time units between baseline harvests
 - t_u = time units between update/addition harvests
 - N_{ttl} = time to live for a posting on a receiving news server
- Sender policies:
- cyclic baseline harvest, no updates between baseline ($t_b > 0$, $t_u = 0$)
 - initial baseline, only updates afterwards ($t_b = 1$, $t_u > 0$)
 - cyclic baseline with updates ($t_b > 0$, $t_u < t_b$)
- Receiver policies:
- $N_{ttl} = 1 \dots \infty$

3.3 SMTP Parameters

- E = attach a record to every "Eth" email
- E_d = # of unique email destinations
- E_r = # of unique email destinations processing attached records ($E_r \leq E_d$)

Sender policies:

- 1 queue: attach to every Eth email regardless of destination
- N queues: attach to every Eth email, remember pointer for each destination (E_d)

Receiver policies:

- N queues with feedback: attach to every Eth email, remember pointer for each destination that told us they are listening (E_r)
- N queues with feedback and updates: adjust E based on R_u & R_a after the destination has received baseline harvest

4. FUTURE WORK

We are in the process of running our simulations and testing the various policies. No results have been published yet, but we expect to submit in Spring 2006.

5. ACKNOWLEDGMENTS

This work supported by NSF ISS 0455997.

6. REFERENCES

- [1] Lagoze, C., Van de Sompel, H., Nelson, M. L., Warner, S. The Open Archives Initiative Protocol for Metadata Harvesting. <http://www.openarchives.org/OAI/openarchivesprotocol.html>
- [2] Nelson, M. L., Van de Sompel, H., Liu, X., Harrison, T. L. and MacFarland,, N. *mod_oai: An Apache Module for Metadata Harvesting*. arXiv.org Technical Report cs.DL/0503069, 2005.

Matching and Integration Across Heterogeneous Data Sources

Patrick Pantel, Andrew Philpot and Eduard Hovy

Information Sciences Institute
University of Southern California
4676 Admiralty Way
Marina del Rey, CA 90292

{pantel,philpot,hovy}@isi.edu

ABSTRACT

A sea of undifferentiated information is forming from the body of data that is collected by people and organizations, across government, for different purposes, at different times, and using different methodologies. The resulting massive data heterogeneity requires automatic methods for data alignment, matching and/or merging. In this poster, we describe two systems, *Guspin*TM and *Sift*TM, for automatically identifying equivalence classes and for aligning data across databases. Our technology, based on principles of information theory, measures the relative importance of data, leveraging them to quantify the similarity between entities. These systems have been applied to solve real problems faced by the Environmental Protection Agency and its counterparts at the state and local government level.

Categories and Subject Descriptors

H.2.5 [Database Management]: Heterogeneous Databases.

General Terms

Algorithms, Experimentation.

Keywords

Information theory, mutual information, database alignment, equivalence class detection.

1. INTRODUCTION

In face of the growing mass of often undifferentiated data being collected in at an unprecedented pace by government agencies, data users need automated assistance to make sense of their data sets by interrelating, clustering, grouping, contained elements etc. The general class of problems is that of finding similarities between entities within or across heterogeneous data sources. To date, most approaches to entity consolidation and cross-source data integration have relied heavily on manual effort and on auxiliary information such as relational structure or metadata. In the work described in this poster, we address the increasingly

common case where metadata is outdated, irrelevant, overly domain specific, or simply non-existent. A general-purpose solution to this problem cannot therefore rely on such auxiliary data. Unfortunately, all one can count on is the data itself: a set of observations describing the entities.

In this “data only” paradigm, we have developed an information theoretic model for matching and integration of data sources. The key to our underlying technology is to identify the most informative observations and then match entities that share them. Applying this model, we have built two systems, *Guspin* for automatically identifying equivalence classes or aliases, and *Sift* for automatically aligning data across databases.

2. INFORMATION MODEL

When interrelating entities based on observational data (e.g., matching people based on their financial transactions and communication patterns), certain observations are much more informative and important and thus indicative of similarity than others. When assessing the similarity between entities, important observations should be weighed higher than less important ones.

Shannon’s classic 1948 paper [4] provides us with a way of measuring the information content of observed events. This theory of information provides a metric, called pointwise mutual information, which quantifies the association between two events by measuring the amount of information one event tells us about the other.

Consider the following scenario, illustrating the power of pointwise mutual information, in which you are a drug trafficking officer charged with tracking two particular individuals *John Doe* and *Alex Forrest* from a population of Southern California residents. If you were told that last year both *John* and *Alex* called Hollywood about 21 times a month, then would this increase your confidence that *John* and *Alex* are the same person or from the same social group? Yes, possibly. Now, suppose we also told you that *John* and *Alex* each called Bogotá about 21 times a month. Intuitively, this observation yields considerably more evidence that *John* and *Alex* are similar, since not many Southern California residents call Bogotá with such frequency. Measuring the relative importance of two such observations—calling Hollywood and calling Bogotá—and leveraging the measurements to compute similarities between entities is the key to our approach.

In our formulation, we use pointwise mutual information to measure the amount of information one event x gives about

another event y , where $P(x)$ denotes the probability that x occurs, and $P(x,y)$ the probability that they both occur:

$$mi(x,y) = \log \frac{P(x,y)}{P(x)P(y)}$$

Given this method of ranking observations by relative importance, we use the cosine coefficient metric [1] to determine the similarity between two entities. In comparison to other candidate metrics, such as Euclidean distance, cosine is less sensitive to *unseen* observations. That is, the absence of a matching observation is not as strong an indicator of dissimilarity as the presence of one is an indicator of similarity. The similarity between each pair of entities e_i and e_j , using the cosine coefficient metric, is given by:

$$sim(e_i, e_j) = \frac{\sum_o mi(e_i, o) \times mi(e_j, o)}{\sqrt{\sum_o mi(e_i, o)^2} \times \sqrt{\sum_o mi(e_j, o)^2}}$$

where o ranges through all possible observations (e.g., phone calls). A similarity of 0 indicates orthogonal vectors (i.e., unrelated entities) whereas a similarity of 1 indicates identical vectors. For two very similar elements, their vectors will be very close and the cosine of their angle will approach 1.

3. SYSTEMS

We have applied this mutual information model to several problems, including automatically building a word thesaurus, discovering concepts, inducing paraphrases, and identifying aliases in a homeland security scenario. In the context of digital government, we have built two web tools, *Guspin* and *Sift*, and applied them to problems faced by the Environmental Protection Agency (EPA). At the core, both systems employ the pointwise mutual information and similarity models described in the previous section.

3.1 Guspin™¹

Guspin is a general purpose tool for finding equivalence classes within a population. It provides a simple user interface where an analyst user uploads one or multiple data files containing observations for a population. The system then identifies and clusters duplicate (or near-duplicate). *Guspin* provides an analyst with a browsing tool for finding equivalence classes and navigating the similarity space of the supplied population. The analyst may also download the resulting *Guspin* analysis for further examination. In an experiment identifying duplicates facilities given between national, state, and local facility catalogs (described in greater detail in our poster), *Guspin* (i) with 100% accuracy, extracted 50% of the matching facilities; (ii) with 90% accuracy, extracted 75% of the matching facilities; (iii) for a given facility and the top-5 mappings returned by the system, with 92% accuracy, extracted 89% of the matching facilities.

3.2 Sift™²

Sift is a web-based application portal for cross-database alignment [2] [3]. Given two relational data sources, *Sift* helps answer the

question “which rows, columns, or tables from data source S_1 have high correspondence with (all or part of) some parallel construct(s) from S_2 ?”. Using domain-independent and domain-dependent probabilistic knowledge-based and syntactic data recognizers (e.g., for phone numbers, CAS registry numbers, SIC/NAICS codes, date/time formats), *Sift* can fortify the as-received observation space with computed observation types. This additional type knowledge is brought to bear during the normal similarity vector space match, allowing for instance that a phone number in S_1 with attached area code might match a phone number in S_2 whose area code is stored in a different column, etc. In an experiment aligning California state and local emissions inventory databases (again described in greater detail on the poster proper) *Sift* discovered 295 alignments, of which 75% were correct. There were 306 true alignments, of which *Sift* identified 221 or 72%. Interestingly, when *Sift* found a correct alignment for a given column, then the alignment appears in the first two returned candidate alignments.

4. CONCLUSIONS

A general-purpose solution to the problem of matching entities within or across heterogeneous data sources cannot rely on the presence or reliability of auxiliary data such as structural information or metadata. Instead, it must leverage the available data (or observations) that describe the entities. Our technology, based on principles of information theory, measures the importance of observations and then leverages these to quantify the similarity between entities. Though the technology is applicable to a wide range of applications, we have built two web solutions, called *Guspin*™ and *Sift*™, addressing the general problems of building equivalence classes or aliases for a population and of aligning heterogeneous databases. These systems have been applied to solve real problems faced by the Environmental Protection Agency and allied state and local environmental quality agencies with remarkable accuracy. Our systems can dramatically reduce the time an analyst needs to find related entities in a population. However, the power of the technology is critically dependent on gathering the right observations that entities might share, which in itself is a very interesting avenue of future work. Our model has the potential to address several serious and urgent problems faced by the government such as terrorist detection, identity theft, and data integration.

5. REFERENCES

- [1] Baeza-Yates, R. and B. Ribeiro-Neto. 1999. *Modern Information Retrieval*. Wokingham: Addison-Wesley.
- [2] Pantel, P.; Philpot, A.; and Hovy, E.H. 2005. Aligning Database Columns using Mutual Information. In *Proceedings of Conference on Digital Government Research (DG.O-05)*. pp. 205-210. Atlanta, GA.
- [3] Pantel, P.; Philpot, A.; and Hovy, E.H. 2005. An Information Theoretic Model for Database Alignment. In *Proceedings of Conference on Scientific and Statistical Database Management (SSDBM-05)*. pp. 14-23. Santa Barbara, CA.
- [4] Shannon, C.E. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal*, 20:50-64.

¹ *Guspin* is available from <http://guspin.isi.edu>.

² *Sift* is available from <http://sift.isi.edu>.

The Impacts of Digital Government on Civic Engagement: A Typology of Information Technology Use

Hun Myoung Park

Indiana University at Bloomington

410 North Park Avenue

Bloomington, IN 47408

812-857-0492

kucc625@indiana.edu

ABSTRACT

This paper proposes a typology to analyze the impact of digital government on civic engagement in politics and policy processes. This framework is based on characteristics of target clients and modes of information technology use (economy, solidarity, and action modes). Major parties and candidates tend to remain in the solidarity model and reinforce advocacy of supporters. In the action model, motivated minority groups and issue-based activists are likely to use one-way and two-way communications effectively and deploy online and offline activities strategically.

Categories and Subject Descriptors

K.4.1 [Computers and Society]: General – *Public participation, civic engagement.*

General Terms

Management, Human Factors, and Theory

Keywords

Public/political participation, civic engagement, economy model, solidarity model, action model, politics on the Internet.

1. INTRODUCTION

Civic engagement refers to citizens' individual and collective involvement in public affairs, which has a wide range of voluntary activities such as voting and attending public meetings [7, 14]. Civic engagement in politics and public affairs lies at the heart of modern democracy. Civic engagement is expected to facilitate good governance and make all levels of government more efficient and accountable [13].

Since the 1980s, information technology (IT) has pervaded the public and private sectors due to its potential to transform the mode of doing business and thus improve organizational productivity. As a variety of IT applications are widely used,

academic attention is given to the impact on various aspects of society. A group of scholars investigates how the use of IT affects public service delivery, government reform, and democracy, while others examine the effects on social capital and communities, digital inequality (digital divide), and organizational and cultural changes [1, 4, 7, 9, 10].

Enthusiasts have a utopian view that IT is able to get people more involved in public life and thus contribute to participatory democracy [5, 9]. The pessimists, however, argue that IT reinforces, rather than transforms, the existing power relationships and patterns of political participation [1, 2, 7]. Finally, the skeptics hold a cautious view that IT, despite its potential, does not necessarily destroy, nor facilitate civic engagement, reflecting the “politics as usual” [6, 12]. This paper is an effort to reconcile these conflicting views by suggesting a typological framework.

2. CLIENTS AND MODES OF IT USE

An abundant literature has endeavored to explore citizen involvement in public affairs [7, 12, 13]. Individual digital government applications affect civic engagement in different ways depending on characteristics of target clients and modes of IT use.

Some IT applications require clients to register first. Registration is required for the sake of efficiency and security in transactions. The second group of applications is open to the homogeneous public who has similar interests. The third type is most widely open even to the heterogeneous public. Anyone comes and goes freely as he/she wants.

The economy mode of using IT applications pursues efficiency; transactions are based on material incentives. In the second mode, users put emphasis more on solidarity and volunteerism rather than efficiency. Finally, the action mode is mainly used by activists who have intensive enthusiasm for their beliefs.

3. TYPOLOGY OF IT USE

Combining characteristics of clients with modes of IT use results in a typology shown in Table 1.

Most e-government applications mainly work in the economy model although advanced e-government is also devoted to reach out to citizens. Current government portals for general public provide a variety of information and links to online services, but are not likely to influence civic engagement substantially. Five percent in 2000 and 29 percent in 2004 of U.S. government portals provide online forums and chat rooms [15].

Table1. Typology of Information Technology Use

	Economy	Solidarity	Action
Membership	G2G, G2B	Minority groups	Exclusive activism
Homogeneous	G2B, G2C	e-campaign, advocacy	Advocacy, cyber-rally
Heterogeneous	G2C (portal)		Strategic use

Major political parties and candidates tend to use IT applications in the solidarity model. Electronic party and e-campaign primarily target party members and supporters who share the political ideology and policy interests. Politicians focus more on reinforcing supporters, fundraising, and mobilizing volunteers rather than persuading adversaries. Candidates try to provide only information that they want constituents to see. The information is displayed in the format they prefer. Politicians tend to eschew two-way channels because of burdensome requirements of these channels, risk of losing control of communications, and loss of ambiguity in campaign discourse [11].

Most citizens take only what they want (narrowcasting); they take a look at “processed” information, watch video clips, download election signs, and contribute money.¹ Despite a slogan of deliberative democracy, only a few citizens post messages, email public officials, or join chats for their business [8]. Not only do candidates try to avoid burdensome interactions, but citizens also do not really demand such two-way communications.

Activists with high degree of enthusiasm resort to the action model. They tend to use IT in order to boost strong ties among members and reach the homogeneous public. When open to heterogeneous public, the action model often suffers from collective action problems, which may disable self-governing and degrade deliberative discussions.² Even if bothered and frustrated, activists do not leave communities easily. Benefits of cost reduction may become less powerful in this model. Activists are not confined to the virtual world, but strategically go back and forth online and offline. This strategic switching may result in bogus online activities that reflect only a part of the real world.

Activists are willing to use both one-way and two-way communications. Poorly resourced advocacy has to look for alternative media such as Internet-based broadcasting and newspapers, since it can not afford to use existing mass media (e.g., TV and radio). As a consequence, the disenfranchised yet self-empowered net activists, who would remain silent without IT, are more likely to use IT applications than major parties and well-established organizations [3]. However, activities of this model are hardly measured in survey research.

4. CONCLUSION

This paper suggests a typology of digital government use that combines the characteristic of clients and the model of Information technology (IT) use. This typological approach provides implications for digital government projects. Online

¹ Online fundraising is the most successful IT application [8].

² A typical example is flaming, the act of posting messages on the Internet that are deliberately hostile and insulting in the social context. <http://en.wikipedia.org/wiki/Flaming>.

forums and chat rooms of government portals need to be managed properly to deal with collective action problems. E-government needs to go beyond the economy model to reach out to citizens. Current e-democracy has limitations in facilitating civic involvement in public affairs.

IT is neither a panacea nor a threat to society. The impacts of digital government on civic engagement depend on how IT applications are properly used in specific circumstances. Thus, it is critical to develop strategies for effective IT use by carefully considering target clients, modes of IT use, and politics on the Internet.

5. REFERENCES

- [1] Bimber, B. The Internet and Political Transformation: Populism, Community, and Accelerated Pluralism. *Polity*, 31, 1 (1998), 133-160.
- [2] Davis, R. *The Web of Politics: The Internet's Impact on the American Political System*. New York: Oxford University Press, 1999.
- [3] Davis, S., Elin, L., and Reher, G. *Click on Democracy: The Internet's Power to Change Political Apathy into Civic Action*. Boulder, CO: Westview Press, 2002.
- [4] DiMaggio, P., Hargittai, E., Neuman, W. R., and Robinson, J. P. Social Implications of the Internet. *Annual Review of Sociology*, 27, 1 (2001), 307-336.
- [5] Grossman, L. K. *The Electronic Republic: Reshaping Democracy in the Information Age*. New York, Viking Penguin. 1995.
- [6] Margolis, M., and Resnick, D. *Politics As Usual: The Cyberspace*. Thousand Oaks, CA, Sage Publications, 2000.
- [7] Norris, P. *The Digital Divide: Civic Engagement, Information Poverty and the Internet Worldwide*. New York, Cambridge University Press, 2001.
- [8] Rainie, L., Cornfield, M. and Horrigan, J. *The Internet and Campaign 2004*. PEW Internet and American Life Project, 2005.
- [9] Rheingold, H. *The Virtual Community*. Addison-Wesley Publishing, 1993.
- [10] Robbin, A., Courtright, C., and Davis, L. ICTs and Political Life. *Annual Review of Information Science and Technology*, 38 (2004), 411-481.
- [11] Stromer-Galley, J. On-Line Interaction and Why Candidates Avoid It. *Journal of Communication*, 50, 4 (Autumn 2000), 111-132.
- [12] Uslaner, E. M. Trust, Civic Engagement, and the Internet. *Political Communication*, 21, 2 (2004), 223-242.
- [13] Verba, S., Schlozman, K., and Brade, H. E. *Voice and Equality: Civic Voluntarism in American Politics*. Cambridge, MA, Harvard University Press, 1995.
- [14] Weissberg, R. *The Limits of Civic Activism: Cautionary Tales on the Use of Politics*. New Brunswick, NJ: Transaction Publishers, 2005.
- [15] West, D. M. E-Government and the Transformation of Service Delivery and Citizen Attitudes. *Public Administration Review*, 64, 1 (Jan./Feb. 2004), 15-27.

Building Regulatory Compliant Storage Systems

Zachary N. J. Peterson

The Johns Hopkins University
3400 N. Charles St.
Baltimore, MD, USA
zachary@cs.jhu.edu

Randal Burns

The Johns Hopkins University
3400 N. Charles St.
Baltimore, MD, USA
randal@cs.jhu.edu

1. PROJECT GOALS

In the past decade, informational records have become entirely digital. These include financial statements, health care records, student records, private consumer information and other sensitive data. Because of the delicate nature of the data these records contain, Congress and the courts have begun to recognize the importance of properly storing and securing electronic records. Examples of legislation include the Health Insurance Portability and Accountability Act (HIPAA) of 1996, the Gramm-Leach-Bliley Act (GLBA) of 1999, and the more recent Federal Information Security Management Act (FISMA) and Sarbanes-Oxley Act (SOX) of 2002. Altogether, there exist over 4,000 acts and regulations that govern digital storage, all with a varying range of requirements for maintaining electronic records.

Some legislation requires that systems provide confidentiality through encrypted storage and data transmission. Some legislation requires an auditable trail of changes made to electronic records that are accessible in real-time. Other legislation sets limits on the amount of time an organization may be liable for maintaining their electronic data. The list of requirements is comprehensive, however, distilling them into product requirements and implementing a system to meet all mandates is non-trivial. It is the goal of this project to make sense of the large body of requirements and develop technical solutions that help organizations manage their data and comply with federal regulations.

2. PROJECT HIGHLIGHTS

We have three major findings to date. We have developed and released an open-source versioning file system designed for regulatory compliance, added secure-deletion to the file system for privacy and compliance with legislation that mandates deletion, and developed a digital audit model that provides a secure record of how data changes over time.

We developed and released the ext3cow versioning file system [9] that addresses the mandated versioning and auditability requirements. The file system provides a *time-shifting* interface that permits a real-time and continuous view of data in the past. Ext3cow has hundreds of users from over one hundred different countries. It has served as tool for other academic research projects, and for us, ext3cow has continued to be a useful foundation for exploring technical solutions to other regulatory storage problems.

While versioning file systems are quickly being adopted by medical and commercial institution wishing to become federally compliant, most existing systems that advertise themselves as “compliant” overlook fine-grained secure deletion

as an essential requirement. Secure deletion is the act of removing digital information from a storage system so that it can never be recovered. Fine-grained refers to removing individual files or versions of file, while preserving all other data in the system. We believe the reticence to integrate secure deletion into existing storage systems derives from the inefficiency and of current deletion techniques when applied to versioning systems. Our second major contribution is the development of two methods for the efficient secure deletion of individual versions of a file that are orders of magnitudes faster than existing techniques [10]. The first method uses all-or-nothing (AON) encryption [11] to create a small *stub* that, when *securely overwritten* [8], permanently deletes the corresponding data. The second technique uses random key generation to generate a stub, similar to *key disposal* [5]. Both techniques provide authenticated encryption [3], which provides both data privacy and authentication. To our knowledge, we are the first disk file system to adopt authenticated encryption. We collect and store stubs contiguously so that overwriting a small block of stubs deletes a large amount of file data, even when file data are non-contiguous. Our methods do not complicate key management.

Another challenge present in compliant storage systems lies in verifying the authenticity of data, *i.e.* making data safe from tampering and providing a proof of compliance. Both auditors and companies are required by SOX to keep strong audit trails on electronic records; for both parties to prove compliance and for auditors to ensure the accuracy of the information on which they report. A “strong” audit trail is a verifiable, persistent record of how and when data have changed. Our third contribution is a system for the verification of version histories in file systems based on generating message authentication codes (MACs) for versions and archiving them with a third party. A file system commits to a version history when it presents the MAC to the third party. At a later time, a version history may be verified by an auditor. The file system is challenged to produce data that matches the MAC, ensuring that the system’s past data have not been altered. The MACs reveal nothing about the data contents and published MACS may even be stored publicly. Our design goals include minimizing the network, computational, and storage resources used in the publication of data and the audit process. To this end, we employ parallel message authentication codes [1, 2, 4] that allow MACs to be computed incrementally – based only on data that have changed from the previous version. Sequences of versions may be verified by computing a MAC for one version

and incrementally updating the MAC for each additional version, performing the minimum amount of I/O. With incremental computation, a natural trade-off exists between the amount of data published and the efficiency of audits.

3. PROJECT STATUS AND ACTIVITIES

Throughout the course of this research, we have been able to effectively collaborate with other researchers at Johns Hopkins and aboard. With secure deletion, we worked with co-PI Avi Rubin and Adam Stubblefield on constructing secure deletion algorithms with authenticated encryption for a version file system. For digital audit trails, we collaborated with co-PI Giuseppe Ateniese and Steve Bono to help in formulating efficient authentication algorithms in a versioning environment.

All projects are being implemented, not simulated, in the ext3cow file system. The ext3cow version file system and an implementation of secure deletion is available for download at www.ext3cow.com. The goal is to provide an open-source implementation of a storage system that meets the requirements of electronic records legislation. This will make compliance available to all, minimizing the costs involved.

We have shared our findings with the public and academic community through publications and conference presentations. Details of the ext3cow file system have been published in the journal, *ACM Transactions on Storage* [9]. Our secure deletion work was most recently published and presented at the USENIX File And Storage Technology (FAST) in December [10]. We have also presented our work on authenticators for versioning file systems at the ACM CCS Workshop on Storage Security and Survivability (StorageSS) [6]. A comprehensive presentation of our research to date was recently made to Digital Archives consortium at the Library of Congress in December of 2005.

4. FUTURE WORK

The development of these tools has left us with many opportunities for continued research in this field. Activities for the next project year include:

- **Secure deletion in managed environments:** We are expanding secure deletion constructs to delete data even when it has been replicated across multiple sites (for backup). We are also developing methods so that users may delete data without physical access to the media, *e.g.* a patient could use this construct to delete portions of her medical records from storage owned by doctors and insurance companies.
- **Unification of secure deletion algorithms:** We are collaborating with Giovanni Di Crescenzo of Telcordia to prove the security of our algorithms for secure deletion and create a unified framework among our work and his work on erasable memories [7].
- **Implementation and release of digital audits:** In the next year, we will complete our implementation of verifiable audit trails in ext3cow and make this available to the public through an open-source license.
- **Approximate MACs:** We have begun an investigation of security constructs that allow for digital audits to be conducted by sampling only portions of the data in a system. Such a construct would greatly improve the efficiency of audits and provide probabilistic guarantees that data have not been modified or lost.

5. REFERENCES

- [1] M. Bellare, O. Goldreich, and S. Goldwasser. Incremental cryptography and application to virus protection. In *Proceedings of the ACM Symposium on the Theory of Computing*, pages 45–56, May-June 1995.
- [2] M. Bellare, R. Guérin, and P. Rogaway. XOR MACs: New methods for message authentication using finite pseudorandom functions. In *Advances in Cryptology - Crypto'95 Proceedings*, volume 963, pages 15–28, 1995. Lecture Notes in Computer Science.
- [3] M. Bellare and C. Namprempre. Authenticated Encryption: Relations among notions and analysis of the generic composition paradigm. In *Advances in Cryptology - Asiacrypt'00 Proceedings*, volume 1976, 2000. Lecture Notes in Computer Science.
- [4] J. Black and P. Rogaway. A block-cipher mode of operation for parallelizable message authentication. In *Advances in Cryptology - Eurocrypt'02 Proceedings*, volume 2332, pages 384 – 397. Springer-Verlag, 2002. Lecture Notes in Computer Science.
- [5] D. Boneh and R. Lipton. A revocable backup system. In *Proceedings of the USENIX Security Symposium*, pages 91–96, July 1996.
- [6] R. Burns, Z. Peterson, G. Ateniese, and S. Bono. Verifiable audit trails for a versioning file system. In *Proceedings of the ACM CCS Workshop on Storage Security and Survivability*, November 2005.
- [7] G. Di Crescenzo, N. Ferguson, R. Impagliazzo, and M. Jakobsson. How to forget a secret. In *Proceedings of the Symposium on Theoretical Aspects of Computer Science*, volume 1563, pages 500–509. Springer-Verlag, 1999. Lecture Notes in Computer Science.
- [8] P. Gutmann. Secure deletion of data from magnetic and solid-state memory. In *Proceedings of the USENIX Security Symposium*, pages 77–90, July 1996.
- [9] Z. Peterson and R. Burns. Ext3cow: A time-shifting file system for regulatory compliance. *ACM Transactions on Storage*, 1(2):190–212, 2005.
- [10] Z. N. J. Peterson, R. Burns, J. Herring, A. Stubblefield, and A. Rubin. Secure deletion for a versioning file system. In *Proceedings of the USENIX Conference on File And Storage Technologies (FAST)*, pages 143–154, December 2005.
- [11] R. L. Rivest. All-or-nothing encryption and the package transform. In *Proceedings of the Fast Software Encryption Conference*, volume 1267, pages 210–218, 1997. Lecture Notes in Computer Science.

Regionalizing Integrated Watershed Management: A Strategic Vision

Keith Pezzoli

Research Scientist

Urban Studies & Planning

University of California San Diego

1-858-534-3691

kpezzoli@ucsd.edu

Richard Marciano

Lead Scientist

San Diego Supercomputer Center

University of California San Diego

1-858-534-8345

marciano@sdsc.edu

John Robertus

Executive Officer

California Regional Water Quality

Control Board San Diego Region

1-858-467-2952

JRobertus@waterboards.ca.gov

ABSTRACT

In addressing the conference theme, *Integrating Information Technology and Social Science Research for Effective Government*, this paper examines the challenges that government agencies face while trying to protect and restore water quality from a watershed management standpoint. Our geographic focus is the San Diego city-region and its neighboring jurisdictions (including Mexico). We find that there is a pressing need to develop a dynamic regional information system that can help guide and track individual development projects (micro-development) in the context of the larger (macro-development) of whole watersheds. Yet serious constraints stand in the way. Fortunately, advances taking place in certain scientific, socio-technical and regulatory domains are promising. Three stand out: (1) the growth of sustainability science and emergence of cyberinfrastructure for multiscale environmental monitoring, (2) the mobilization of what the National Research Council calls *knowledge-action collaboratives*—including university-government-community partnerships, and (3) regulatory innovation calling for watershed-based approaches to environmental policy and planning. We need a concerted strategy to integrate and take full advantage of these trends. This paper provides a strategic vision along such lines. A case study on digital systems for environmental mitigation and tracking is also presented. Digital government research themes related to this case study include: (1) long-term preservation and archiving of government records, (2) integration of data grids and geographic information systems, and (3) citizen interactions through transparency of and universal access to digital records.

Keywords

Digital government, Integrated watershed management, Cyberinfrastructure, Sustainability science, Long-term preservation, Data grids, Citizen interaction.

1. INTRODUCTION

This paper examines the challenges that government agencies face while trying to protect and restore water quality from a watershed management standpoint. Our geographic focus is the San Diego city-region and its neighboring jurisdictions (including Mexico). We find that there is a pressing need to develop a dynamic regional information system that can help guide and track individual development projects (micro-development) in the context of the larger (macro-development) of whole watersheds. This paper examines the constraints and opportunities to building such a system.

2. THE SAN DIEGO CASE

Key San Diego government agencies such as the San Diego Regional Water Quality Control Board (SDRWQCB, known as the “Regional Board”) are increasingly convinced that integrating various watershed data and trends at a regional scale is essential. Nine Regional Boards cover the entire state of California. The overarching mission of these agencies is “to preserve, enhance and restore the quality of California’s water resources, and ensure their proper allocation and efficient use for the benefit of present and future generations” (State Water Resources Control Board Mission Statement). The SDRWQCB, and other Regional Boards throughout California, have begun grappling with issues of integrated watershed management.

3. THREE ADVANCES

3.1 Sustainability and Cyberinfrastructure

The growth of sustainability science and emergence of cyberinfrastructure for multiscale environmental monitoring.

OBSTACLES: We still have a seriously fragmented knowledge base (disciplinary boundaries create knowledge silos that thwart science-science integration); no mandate to build regional cyberinfrastructure ---it happens on an ad hoc and duplicative basis given the balkanized and competitive landscape of grant seeking.

SOLUTIONS: Our approach was to build a RWBC/SALT initiative, building social capital through university-community-government partnerships. sustainability science / cyberinfrastructure consortium, based on the Regional Workbench Consortium (RWBC), a "knowledge-action collaborative" geared to linking science and technology to policy and planning for sustainable city-region development [1]. Funded in large part by the Outreach Core of UCSD's Superfund Basic Research Program since 2000, the RWBC takes a forward-looking perspective by focusing on the Southern California-Northern Baja California transborder region - especially the San Diego-Tijuana city-region and coastal zone. The SALT laboratory (Sustainable Archives & Library Technologies) at the San Diego Supercomputer has teamed up with the RWBC to advance the notion of “the digital watershed” by way of a prototype called **WET** (Watershed Exploration Tool), which integrates document management systems, databases, and GIS and operates in a data grid environment (a framework which allows for distributed management of data collections)..

3.2 Knowledge-Action Collaboratives

The mobilization of what the National Research Council calls knowledge-action collaboratives—including university-government-community partnerships.

OBSTACLES: We need much more in the way of university-led translational research (i.e., the sharing of basic research with prospective end users), but the incentive system for this is weak. Partnerships linking science-to-society are difficult to establish and maintain over the long haul.

SOLUTIONS: The rising demand for “accountability” in federal research grants creates opportunities. Many larger federal grant programs now routinely require efforts to identify the larger societal benefits of basic research. For instance, the NSF awards research funding based on merits of the proposed activities that include “their potential impact on society.” The UCSD Superfund Basic Research Program (2005-2010) [2], funded by a five-year grant from the National Institute of Environmental Health Sciences (NIEHS) now has a mandated “Research Translation Core.” This emphasis on accountability and benefits for the larger common good is promising. It provides incentives for academics to collaborate in linking science to society ---what some have referred to as building Ivory Bridges.

3.3 Watershed-Based Approaches

Regulatory innovation calling for watershed-based approaches to environmental policy and planning.

OBSTACLES: Environmental policy is still largely driven by a “command and control” paradigm that is highly problematic for many reasons.

SOLUTIONS: The formation of watershed partnerships has emerged as a favored strategy to improve regional economic and regulatory efficiencies in environmental management, especially water pollution prevention (P2). Also, watershed-based frameworks that could lead to integrated information technology systems, are being refined. Co-author John Robertus, is compiling a macro-development approach which focuses the watershed and regional scope of concern [3]. The task is to reasonably determine the relative status of the water quality impacts from previous development in the watershed so that each project can be accurately evaluated and adequately mitigated given the true water quality conditions in the watershed. The Macro-Development water quality status for a watershed includes many factors including the following:

1. The extent of impacts from water polluting discharges from existing development.
2. The impacts of hydromodification to the original creeks, streams, rivers, lagoons and bays
3. The amount and location of watershed area that has been covered with impervious surfaces
4. The success of economic and political efforts to protect water quality in the watershed

One of the most tangible outcomes of working towards understanding and building the foundation of a watershed-based planning support system is the ability to seamlessly integrate

information from multiple spatial sources including a mitigation / Best Management Practices database, economic indicators, GPS citizen-collected data, EPA-maintained databases such as TRI, publicly available internet map servers.

4. CASE STUDY: A WATERSHED TOOL

Our approach is based on creating a detailed historical record of the health of watersheds and rivers and the fact that the most needed commodity is accurate information for all parties involved. Credibility of content should not be dependent on the creator of the data. The prototype we built is called WET or Watershed Exploration Tool.

The data collected and organized is managed by a persistent archive at the San Diego Supercomputer Center, based on the use of the Storage Resource Broker (SRB) data grid. This technology is being developed as part of the Data Intensive Computing Environments group. The use of underlying data grid technology allows us to consider watershed-based infrastructure, where we can scale the approach to other watersheds, manage the data in a distributed fashion, created shared collections, accessible by various stakeholders. This kind of approach is already under way and being tested in the context of the Persistent Archives Testbed (PAT) project at the San Diego Supercomputer Center, where we built a “community” grid with linked storage resources located remotely at the State of Minnesota, Michigan, Kentucky, Ohio, Stanford, and San Diego Supercomputer Center [4].

The project experiments with portal design that provides multiple linked interfaces: querying of the data itself, of the digitized documents and of the maps, thus providing a unified perspective across projects, geography and time.

The other layers convey this notion of the “new archives”, where content is made useful to a variety of audiences and viewpoints. One of the main features illustrated in the prototype was the integration of a document-approach and a GIS-approach. This allows the user to select and navigate through a spatial interface, while jumping over at any point to a document interface.

5. ACKNOWLEDGMENTS

This research was partly supported by the NARA/NHPRC Persistent Archive Testbed (PAT) project grant and the National Institute of Environmental Health Sciences, Superfund Basic Research Program, UC San Diego.

6. REFERENCES

- [1] Regional Workbench Consortium, RWBC, <http://regionalworkbench.org>
- [2] USCD Superfund Basic Research Program (2005-2010), <http://superfund.ucsd.edu>
- [3] Robertus, J. CLE 26-27 Jan 06 -- *Water Quality Regulatory Dynamics of Development* -- CLE International California Wetlands Conference -- January 26-27, 2006 -- San Diego, California
- [4] Persistent Archives Testbed (PAT) project, <http://www.sdsc.edu/PAT>.

Distributed Higher-Order Text Mining: Theory and Practice

William M. Pottenger

CSE Department, Lehigh University

billp@lehigh.edu

Shenzi Li

CSE Department, Lehigh University

shl3@lehigh.edu

Christopher D. Janneck

CSE Department, Lehigh University

cdj2@lehigh.edu

ABSTRACT

This highlight discusses the current and ongoing research into distributed higher-order text mining, as implemented using the DiHO ARM algorithm in the D-HOTM system. The DiHO ARM algorithm performs association rule mining in the absence of full knowledge of a global schema on distributed data that is neither vertically nor horizontally fragmented. The D-HOTM system encapsulates the DiHO (and potentially any other) rule-mining algorithm in a distributed system, designed as an extensible digital toolset for data analysts in law enforcement, counterterrorism, health care and other application domains.

Categories and Subject Descriptors

H2.8 [Database Management]: Database Applications - *data mining*, H.2.4 [Database Management]: Systems - *distributed databases, textual databases*, J.1 [Computer Applications]: Administrative Data Processing - government.

General Terms

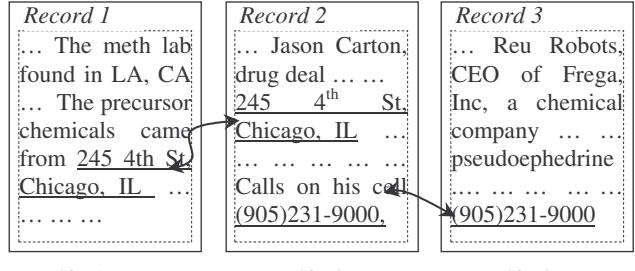
Algorithms, Theory, Design, Performance, Experimentation.

Keywords

Association Rule Mining, Distributed Data Mining, Ensemble Learning, Machine Learning, Data Mining, Text Mining, Text Analytics, Information Extraction, Distributed ARM

1. INTRODUCTION

The spread of information technology and subsequent accumulation of data has brought mining technology to the forefront of analytical toolsets. Association Rule Mining (ARM) is a well known technique that focuses on item co-occurrence within the context of a transaction or record. While most ARM work actively utilizes 1st-order co-occurrence (itemset creation within a single record), this work extends ARM to discover higher-order co-occurrences (itemset creation both within and between records) in distributed environments. This extended approach can be applied to many domains, including law enforcement and homeland defense. An example of a higher-order (3rd-order) link is in Figure 1. This link would result in an association rule such as: *meth lab* \Rightarrow *Reu Robots*.



Site1 Site2 Site3
Figure 1. An Example of Higher-Order Association

2. DiHO ARM ALGORITHM

The Distributed Higher-Order Association Rule Mining (DiHO ARM) algorithm is designed to discover such higher-order associations across distributed data sources. This algorithm is at the forefront of link analysis technology as it overcomes several restrictions often imposed on mining algorithms, namely: higher-order ARM in a distributed environment (other algorithms are local); rule discovery across hybrid (neither horizontally nor vertically) fragmented data (other algorithms are constrained to only one type of fragmented data); and does not require full knowledge or establishment of a complete global schema (others do). DiHO ARM is outlined in Figure 2, with a more detailed description given in Li, Pottenger and Wu (2005).

- (1) Select linkage items
- (2) Assign a globally unique ID to each record/object
- (3) Apply Apriori to the globally unique object IDs
- (4) Identify linkable records
- (5) Exchange information about linkable records
- (6) For each site:
- (7) Apply the Apriori algorithm locally

Figure 2. The DiHO ARM Algorithm

3. D-HOTM THEORY AND SYSTEM

3.1 Theoretical Framework

The D-HOTM framework is designed to discover rules based on higher-order associations in a complex distributed environment. In Li et al. (2006), we introduce the theoretical foundations for reasoning about record linkage assuming no errors in record linkage. This foundation rests on several theorems presented in Li et al. (2006). Following this in the same paper, a sketch of the theoretical framework needed to incorporate errors in record linkage is presented. This latter work is essential to incorporate the traditional support and confidence evaluation metrics into a distributed ARM algorithm.

3.2 D-HOTM Architecture

The Distributed Higher-Order Text Mining (D-HOTM) system is designed to be a flexible, extendable research platform for distributed data mining research and development. It is based on the Text Mining Infrastructure (Holzman et al., 2004), a reusable

framework conducive to conducting research and development in text/data mining. As such, the D-HOTM design comprises three major loosely-coupled components: a control component, a D-HOTM core component, and an analysis component. The control component allows the user to specify all parameters regarding the operation of the mining job – threshold values, output types, algorithms to run, etc. The D-HOTM core is composed of the DiHO ARM and related algorithms. This core uses the parameters specified by the control component and sends the generated output to the analysis component. The analysis component contains a suite of tools allowing users to parse, filter, graph, examine and explore the data provided by the mining process.

3.3 D-HOTM Core Architecture

As noted, the D-HOTM core component is comprised of the central mining algorithms, and performs the bulk of the computational work. This core provides for multiple algorithms to be used in this system, and to date a preliminary prototype of the DiHO ARM algorithm has been implemented and tested. Surrounding the algorithm core are managers for data input (Entity Extractors), output (Rule Managers), and a checkpointing system, allowing for mid-processing states to be saved and restored as a safeguard against hardware failure or to be studied in a post-processing phase. The DiHO ARM implementation also consists of several key classes that relate to the steps of the algorithm, as shown in Figure 3.

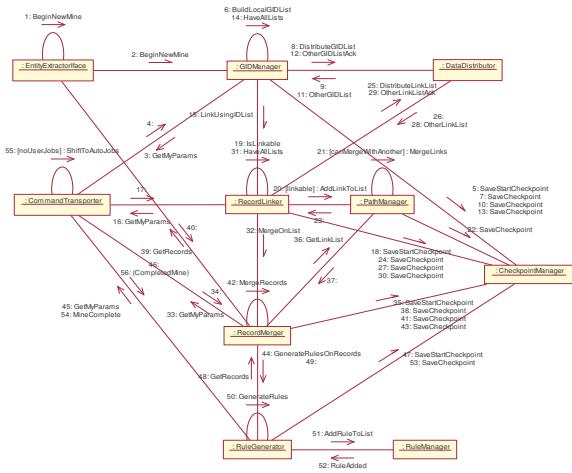


Figure 3. Collaboration Diagram for DiHO ARM

3.4 Implementation and Testing

At the time of writing, as noted a prototype of the D-HOTM core has been implemented and tested on the National Center for Supercomputer Applications' (NCSA) Tungsten Supercluster (Xeon Linux). Using a simulated dataset of methamphetamine trafficking processes, comprising of 4,000 records and utilizing 128 processes on 64 nodes, the D-HOTM prototype generates both first and higher-order rules in less than 18 seconds. The generated rules have also been validated for correctness, per the design of the input data. As noted the system is based on the TMI, and utilizes the Message Passing Interface (MPI) for

distributed inter-process communication. Further details on the implementation and experiments can be found in Li et al. (2006).

4. CONCLUSION AND FUTURE WORK

We have embarked on an ambitious program of research and development that addresses significant challenges in distributed data management faced by organizations such as law enforcement agencies and healthcare providers. We have identified critical assumptions made in existing association rule mining algorithms that prevent them from scaling to complex distributed environments in which the complete global schema is unknown, data is fragmented in a hybrid non-vertical, non-horizontal form, and errors occur in record linkage. We developed a theoretical framework to reason about record linkage, and a theoretical framework for evaluation metrics based on linkage matching errors. We also designed, implemented and tested a prototype of a distributed higher-order association rule mining algorithm, DiHO ARM, which discovers propositional rules based on higher-order associations in a distributed environment.

In our future work we plan to address both theoretical and practical issues in areas such as further exploration of the utility of higher-order associations as well as record linkage, evaluation metrics and issues in efficiency of execution. In addition, a GUI is being developed using Java SWING that will receive and transmit the parameters of a mining job (and mid-processing controls such as pausing or killing) to the distributed D-HOTM core. Also, work is being done to determine what visualization and computational steering methods can be applied in D-HOTM. Finally, by conforming to industry standards, we plan to develop and deploy a toolset that will be quickly, readily, freely and effectively applied by law enforcement personnel in the fight against terrorism.

The TMI, D-HOTM system, and future related research projects are or will be available at <http://hddi.cse.lehigh.edu>.

5. ACKNOWLEDGMENTS

This work was supported in part by NSF grant number 0534276 and NIJ grant numbers 2003-IJ-CX-K003, 2005-93045-PA-IJ and 2005-93046-PA-IJ. The authors also wish to thank Jesus the Messiah for His work of salvation in their lives.

6. REFERENCES

- [1] Holzman, L. E., Fisher, T. A., Galitsky, L. M., Kontostathis, A. and Pottenger, W. M. (2004) A Software Infrastructure for Research in Textual Data Mining. *The International Journal on Artificial Intelligence Tools*, Volume 14, Number 4, Pages 829-849.
- [2] Li, S., Wu, T. and Pottenger, W. M. (2005) Distributed Higher Order Association Rule Mining Using Information Extracted from Textual Data. *SIGKDD Explorations*, Volume 7, Issue 1, June.
- [3] Li, S., Janneck, C. D., Pottenger, W. M., Belapurkar, A. P., Ganiz, M. and Wu, T. (2006) Mining Higher-Order Association Rules from Distributed Named Entity Databases. Submitted to *KDD '06*.

Scalable and Secure Data Collection: Incentives Considerations*

Ranjit Raveendran
Computer Science
Department
University of Southern
California
Los Angeles, CA
raveindr@usc.edu

William C. Cheng
Computer Science
Department
University of Southern
California
Los Angeles, CA
bill.cheng@acm.org

Leana Golubchik
Computer Science and
EE-Systems Departments,
IMSC, and ISI
University of Southern
California
Los Angeles, CA
leana@cs.usc.edu

ABSTRACT

Data collection, or *uploading*, is an inherent part of numerous digital government applications. In this poster we present our recent research directions in the development of Bistro, a scalable and secure architecture designed for collection of data over the Internet for digital government applications. Specifically, we focus on consideration of incentives for participation in the Bistro infrastructure.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Retrieval models; H.3.4 [Systems and Storage]: Distributed Systems; C.4 [Performance of Systems]: Reliability, availability, and serviceability; J.4 [Social and Behavioral Sciences]: Economics

General Terms

Design Performance

Keywords

Incentives, Trustworthiness, Scalable data collection

1. INTRODUCTION

Hotspots are a major obstacle to achieving scalability in the Internet; they are usually caused by either high demand for some data or high demand for a certain service. At the application layer, hotspot problems have traditionally been dealt with using some combination of increasing capacity, spreading the load over time and/or space, and changing the workload. These classes of solutions have been studied in the context of applications using the following types of communication: (a) one-to-many (data travels primarily from a

server to multiple clients, e.g., web download, software distribution, video-on-demand); (b) many-to-many (data travels between multiple clients, through either a centralized or a distributed server, e.g., chat rooms, video conferencing); and (c) one-to-one (data travels between two clients, e.g., e-mail, e-talk). However, to the best of our knowledge ours is the first work on making applications using *many-to-one* communication scalable and efficient; existing solutions, such as web based uploads, simply use many independent one-to-one transfers. This corresponds to an important class of applications, whose examples include a large number digital government applications.

Specifically, government at all levels is a major *collector* and provider of data, and there are clear benefits to disseminating and collecting data over the Internet, given its existing large-scale infrastructure and wide-spread reach in commercial, private, and government domains. In this project, we focus on the *collection of data over the Internet*, and specifically, on the *scalability* issues which arise in the context of Internet-based massive data collection applications. By data collection, we mean applications such as Internal Revenue Service (IRS) applications with respect to electronic submission of income tax forms. The Integrated Justice Information Technology Initiative facilitates information sharing among state, local, and tribal justice components. An integrated (global) information sharing system involves collection, analysis, and dissemination of criminal data. In order to facilitate such a system one must provide a *scalable* infrastructure for collection of data. Briefly other such applications are as follows. A number of government agencies (e.g., NSF, NIH) support research activities, where the funds are awarded through a grant proposal process, with deadlines imposed on submission dates. The entire process involves not only submission of proposals, which can involve fairly large data sizes, but also a review process, a reporting process (after the grant is awarded), and possibly a results dissemination process. All these processes involve a data collection step. Digital democracy applications, such as online voting during federal, state, or local elections, constitute another set of massive upload applications. There are numerous other examples of digital government applications with large-scale data collection needs.

Our past work proposed Bistro [1, 2], a framework for building scalable and secure wide-area digital government *upload*

*This work is supported by the NSF Digital Government Grant 0091474. More information about the Bistro project can be found at <http://bourbon.usc.edu/iml/bistro>.

applications. Briefly, the Bistro upload architecture works as follows. Given a large number of clients that need to upload their data by a given deadline to a given destination server the Bistro architecture breaks the upload problem into three steps:

- *Step 1*, the timestamp step, which must be accomplished prior to the deadline for clients to submit their data to the destination server, also known as destination bistro (DB). In this step, each client sends to the server a message digest of their data and in return receives a secure timestamp ticket from the destination server as a receipt indicating that the client made the deadline for data submission. The purpose of this step is to ensure that the client makes the deadline *without* having to transfer their data which is significantly larger than a message digest and might take a long time to transfer during high loads which are bound to occur around the deadline time. It is also intended to ensure that the client (or an intermediate bistro used below) does not change their data after receiving the timestamp ticket (hence the sending of the message digest to the destination server). All other steps can occur before or after the deadline.
- *Step 2*, the transfer of data from clients to intermediate bistros (IBs). This results in a low data transfer response time for clients since (a) the load of many clients is distributed among multiple bistros and (b) a good or near-by bistro can be selected for each client to improve data transfer performance. Since the bistros are not trusted entities (unlike the DB), the data is encrypted by the client prior to the transfer.
- *Step 3*, the collection of data by the DB. The destination server determines when and how the data is collected in order to avoid hotspots around the destination server. Once the destination server collects all the data, it can decrypt it, recompute message digests, and verify that no changes were made to a client's data (either by the client or by one of the intermediate bistros) after the timestamp ticket was issued.

A summary of main advantages of the Bistro architecture is as follows: (1) hotspots can be eliminated around the server because the transfer of data is decoupled from making of the deadline, (2) clients can receive good performance since they can be dispersed among many bistros and each one can be direct to the "best" bistro for that client, and (3) the destination server can minimize the amount of time it takes to collect all the data since now it is in control of when and how to do it (i.e., Bistro employs a server pull).

Our past work focused on performance, security, and fault tolerance issues. This poster briefly outlines new research directions on incentives for participating in the Bistro infrastructure.

2. INCENTIVES CONSIDERATIONS

We envision a common Bistro infrastructure in which multiple destination servers, also called destination bistros (DBs), corresponding to different applications (e.g., submissions of

tax forms to IRS or proposals to NSF) would compete for a common set of resources, which are termed as Intermediate Bistros (IBs). The IBs are public hosts in the Bistro infrastructure which provide temporary storage and facilitate significantly higher data collection performance.

The focus of this research direction is on designing incentive schemes for encouraging (non-malicious and reliable) participation in the infrastructure. We are currently pursuing a reputation based approach to this problem.

Reputation is a measure of how trustworthy a bistro has been in the past. It is also indicative of how much of its own resources a bistro had contributed to aiding others in the infrastructure. The higher the reputation of a bistro, the higher preference it would receive in the allocation of the infrastructure's resources.

The incentive schemes are needed to encourage bistros to volunteer their resources as well as to incentivize nodes that are currently contributing resources to behave in a reliable and non-malicious manner. (Examples of malicious behavior include corruption of data or reluctance to forward data to an appropriate destination).

While designing an allocation mechanism that considers the reputation of the DBs, the fairness of the resulting allocation also matters. Being fair should encompass the fact that there could be multiple types of resources (e.g. some bistros may be more reliable than others or may have better performance). Developing such an allocation mechanism that optimizes fairness over various types of resources is a difficult problem. This difficulty is further compounded by the fact that the participating DBs (a) could value the same combination of resources differently and (b) could request different combinations of resources.

In summary, this research direction presents a number of algorithmic and distributed system design challenges.

3. CONCLUSIONS

Our work thus far indicates that the goal of achieving an efficient, scalable, secure, and fault-tolerant means for data collection is possible over the public Internet for digital government applications. We believe that the solution of the incentives problem raised in this poster will result in further significant performance improvements for such applications.

4. REFERENCES

- [1] S. Bhattacharjee, W. C. Cheng, C.-F. Chou, L. Golubchik, and S. Khuller. Bistro: a platform for building scalable wide-area upload applications. *ACM SIGMETRICS Performance Evaluation Review (also presented at the Workshop on Performance and Architecture of Web Servers (PAWS) in June 2000)*, 28(2):29–35, September 2000.
- [2] W. Cheng, C. Chou, L. Golubchik, S. Khuller, and H. Samet. Scalable data collection for internet-based digital government applications. In *1st National Conference on Digital Government Research*, pages 108–113, Los Angeles, CA, May 2001.

Elements of Social Science Engagement in Information Infrastructure Design

David Ribes¹, Karen S. Baker²

¹ Sociology/Science Studies
University of California, San Diego
La Jolla, CA 92093, USA
1.858.534.4627
weber.ucsd.edu/~dribes
dribes@ucsd.edu

²Scripps Institution of Oceanography
University of California, San Diego
La Jolla, CA 92093-0218, USA
1.858.534.2350
baker@ucsd.edu

ABSTRACT

Drawing on three cases of information infrastructure building projects with social science participants, we identify four elements which have structured the engagements. The elements we identify are (i) the temporal initiation of social science engagement with the project; (ii) the level of development of the infrastructure at engagement, (iii) the project's participatory model for social science, and; (iv) social scientist's structured relations to project participants.

Categories and Subject Descriptors

K.4.3 [Organizational Impacts]: Computer-supported collaborative work

General Terms

Management, Design, Human Factors

Keywords

data interoperability, infrastructure, collaboration methodology, social science, organization, CSCW, science and technology studies, ethnography

1. OVERVIEW

A new space for social science is opening within large-scale technical projects with associated information infrastructure building, frequently designated cyberinfrastructure. These endeavors are framed as complex and ambitious combinations of information technology enactment, research goals, knowledge outputs, and the bringing together of diverse communities. This framing represents an opportunity for the participation of social scientists not only as researchers, but also as project participants in the creation of collaboratories, standards, metadata languages, ontologies, ‘best practices,’ and other design and implementation work. However, within these projects ‘intervention’, ‘collaboration’ and ‘participation’ – the engagement and contribution of the analyst to the field of action – itself remains an under-explored topic. In these examples ‘the field of action’ is the collaborative practice of science with its associated multiple discipline team dynamics. In this poster we analyze social science engagements with three scientific

information infrastructure projects, drawing from fields of social informatics [4], sociotechnical organizational theory [2]. Our goal is to begin cataloguing the properties of structured relationships between infrastructure projects, social science collaborators, and avenues for intervention. In turn, we intend this research to inform the design and planning of future infrastructure endeavors.

Our three sites of investigation are part of a larger comparative project of the strategies for developing interoperability within large-scale and long-term science (cyber)infrastructures [5]. The three research cases are: GEON, an umbrella cyberinfrastructure for the earth-sciences; the Long Term Ecological Research Program; and Ocean Informatics. Having a research agenda that is strongly coupled to a goal of contributing back to these projects has provided insights into the multiple strategies of collaborative work.

In this poster we show how the abilities of social scientists to contribute within large-scale technical projects has been substantially structured by the configuration of relationships established with the infrastructure projects. In each of GEON, LTER and OI we are participants; but our engagement with each project, and the nature of each project itself, differs in terms of access to the research site, when and how we became involved and what venues exist for communicating findings or collaborating in design.

2. MODELS OF SOCIAL SCIENCE ENGAGEMENT

Differences in style of intervention emerge both from within the nature of the infrastructure projects themselves and from the specificities of the engagement with social science researchers. In making a structural analysis of the positions of the social scientist within engagements, we have consider four elements:

- (i) the *temporal initiation* of social science engagement with the project,
- (ii) the *level of development* of the infrastructure at engagement,
- (iii) the *participatory model* of social science in the project, and;
- (iv) social scientist's *structured relations* to other project participants.

		GEON	OI	LTER
Infrastructure Project	Social Science Temporal Initiation	At formal inception: i - post-proposal; ii -post- funding; iii -pre-enactment.	At planning stage: i- pre-proposal; ii - pre-funding; iii- pre-enactment.	At maturity: i- post- proposal; ii- post-funding; iii- post-enactment.
	Level of Deployment At Social Science Engagement	Written and funded proposal, no organizational or technical structure enacted.	Unfunded proposal. No institutional recognition. Technical infrastructure coordination in progress.	Well established organization, communication and technological infrastructure.
Community Engagement	Participatory Model	Social Dimensions Feedback	Collaborative Design	Network Propagation (local collaborative design)
	Relation of Social Scientists to Infrastructure Project	Observation; Project requested presentations at collective meetings; Informal feedback	Stakeholders; Embedded; Participation in multiple design teams and community events	Stakeholders; Single site findings propagated across existing communication network

Table 1. Elements of Social Science Engagements

We have divided these elements for analytic purposes. However, to understand the positions of the social scientist in the structure of the engagement these elements must be understood in combination.

We can roughly describe (i) and (ii) as development features of the infrastructure project. Thus within GEON the social science relationship began at formal inception. At this point the level of development of the infrastructure was ‘made of’ conceptual plans as outlined in the written proposal and shared by project PI’s. These PI’s had begun to form a social network across multiple institutions (such as the SDSC, UTEP, and Virginia Tech), and had secured a promise of finances from the NSF. At formal inception, then, GEON already had a certain trajectory. In contrast the social science engagement with OI is most accurately described as beginning *before* OI. At this point the level of the development of the infrastructure is ‘made of’ informal social networks from which, over time, proposal writing and other collaborative activities began to produce a vision for OI. Within LTER the engagement began with a matured and highly structured organization maintaining a developed vision, technical infrastructure and multiple means of communication and organization. This leads directly to (iii) and (iv) which we can roughly describe as the organizational aspects of the social science engagement. Both OI and LTER engagements have been informed by similar participatory models which we call collaborative design [1, 3]. LTER and OI projects differ by the availability of networks developed for propagation. In contrast, the GEON engagement has been developed with a participatory model we call ‘social dimensions feedback’, in which the structured relations with social scientists can be characterized by project requested presentations at collective meetings on topics such as communication, culture and community. Table 1 summarizes the ties between the four elements in the three infrastructure projects we cover in the poster. The table suggests the kinds of possible interventions that emerge at the intersection of multiple elements.

We take configurations of social science engagements within infrastructure projects as themselves constitutive of varying spaces for purposeful action. The outcome of this research is **not**

a defined list of the ‘four key points to the social dimensions of infrastructure.’ Rather, we seek to produce the resources by which fruitful relations between social and information sciences may be designed and which will then in turn be capable of successfully addressing local and emergent needs within infrastructure projects. We argue that the varying combinations of these elements substantially inform the possibilities for social science contributions to each project. In considering future projects with social science collaboration within infrastructure building projects, careful consideration of these elements will serve to provide insight and facilitate design of the collaborative engagement.

3. ACKNOWLEDGEMENTS

This work supported by NSF grants including HSD #04-33369 and SES #05-25985. We would like to thank our colleagues Geoffrey C. Bowker and Florence Millerand.

4. REFERENCES

- [1] Baker, K.S., Ecological Design: An Interdisciplinary, Interactive Participation Process in an Information Environment. in *Proceedings of the Workshop on Requirements Capture for Collaboration in e-Science, January 14-15*, (Edinburgh, 2004), 5-7.
- [2] Fountain, J.E. *Building the Virtual State: Information Technology and Institutional Change*. Brookings Institution Press, Washington, D.C., 2001.
- [3] Jackson, S.J. and Baker, K.S. Ecological Design, Collaborative Care, and Ocean Informatics. *Proceedings of the Participatory Design Conference, Toronto*.
- [4] Kling, R. What is Social Informatics and Why Does it Matter? *D-Lib Magazine*, 5 (1). 1-23.
- [5] Ribes, D., Baker, K.S., Millerand, F. and Bowker, G.C. Comparative Interoperability Project: Configurations of Community, Technology, Organization. *Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries*.

Effective Citizen Relationship Management: Hurricane Wilma and Miami-Dade County 311

Alexander Schellong
Harvard University

Johann Wolfgang Goethe- University, Frankfurt a. M.
alexander_schellong@ksg.harvard.edu
schellong@em.uni-frankfurt.de

Thomas Langenberg
Harvard University

Ecole Polytechnique Fédérale de Lausanne (EPFL)
thomas.langenberg@ksg.harvard.edu
thomas.langenberg@epfl.ch

ABSTRACT

As recent events have shown, effective knowledge sharing has become important at all political levels, especially when disasters occur. In this poster, we present the case of Miami-Dade County, which implemented a multi-jurisdictional, multi-channel environment (311/portal) and successfully utilized it during Hurricane Wilma. Drawing from our research on citizen relationship management (CiRM) and literature on absorptive capacity (ACAP), we argue that this setting increase an organization's ability to acquire, assimilate, transform, and exploit information and knowledge regarding the citizen's needs. We shall discuss implications for further CiRM research and managerial insight for emergency management.

General Terms

Citizen Relationship Management, absorptive capacity, emergency management, 311

Keywords

Customer Service, customer relationship management, knowledge management, disaster, emergency management, CRM, ACAP, call center

1. INTRODUCTION

Natural disasters usually have dramatic effects on wide geographic areas, millions of people and are complex to manage. Successful disaster management strongly depends on local, state, and federal government's capacity to deal with such events. Capacity in this context refers to sufficient funding, effective communication, operation procedures/emergency plans, training, public education, and collaboration across state levels and agency boundaries.

Citizen Relationship Management (CiRM) draws from the commonly known concept of Customer Relationship Management (CRM) in the private sector, and is a cluster of management practices, communication channels and technological solutions to handle issues, problems, concerns, and demands of the citizen. Among the goals of CiRM is improving

citizen orientation, enhancing accountability and changing the citizen government relationship. CiRM is a relatively new sub-field of scholarly research, and thus provides considerable potential for both theoretic and empirical research. In current public sector practice, call centers (especially "311") and web-centric citizen service centers (such as one-stop-shop government portals) represent the most common forms of CiRM projects. At the moment, cross agency and cross jurisdictional collaboration exists only at the technology level. The organizational value of CiRM, such as effective knowledge management and knowledge exploitation, has not been fully explored yet.

In the management literature, scholars developed the absorptive capacity (ACAP) construct to clarify why organizations differ in their ability to acquire, assimilate, transform, and exploit external knowledge sources and experience. Researchers use the ACAP construct to explain how individuals within organization seek novel information, build information relationships, or transfer knowledge across organizational boundaries.

In this poster, we draw from ACAP literature and argue that ACAP presents a worthwhile perspective to look at emergency management. We hypothesize that CiRM increases a public administration's (PA) absorptive capacity, by providing it with the technical and organizational means to effectively manage information flows during emergency situations. We test our hypothesis with cross-sectional data from Miami-Dade County's 311 system and its utilization during the Hurricane Wilma event from 10/20 through 11/1/2005. With the poster we attempt to make two contributions. First, we illuminate the ACAP perspective as another important dimension of effective CiRM. Second, we draw managerial insights from the case study and discuss their implications for effective emergency management.

2 Theoretical Background

2.1 Citizen Relationship Management

The term Citizen or Constituent Relationship Management is derived from Customer Relationship Management. CRM typically builds on information

technology and a variety of channels to interact with customers. Successful CRM furthermore requires a customer centric business philosophy, effective business processes, and often dramatic cultural and organizational change at the business firm.

2.2 Emergency Management

In the literature, scholars identify four distinct phases of emergency management: preparedness, response, recovery and mitigation [2]. CiRM has important implications for effective emergency management. The ability of public administration and emergency management offices to provide real-time information and communication regarding the status quo determines the success of the overall emergency management.

2.3 Absorptive Capacity

Effective knowledge sharing and learning is an important source of organizational performance. Ever since Cohen and Levinthal [1] introduced the absorptive capacity construct (ACAP), scholars frequently used it to explain why organizations vary in their ability to either profit from or exploit novel knowledge. So far, scholars used the ACAP construct to account for a couple of phenomena. One line of research argues that ACAP explains how organizations acquire and assimilate novel knowledge from external sources. Another line of research focuses on ACAP as a predictor of the organization's ability to transform internalized knowledge in order to exploit and subsequently increase organizational performance with it. The four key dimensions of ACAP are knowledge acquisition, assimilation, transformation and exploitation.

In summary, why is it worthwhile looking at absorptive capacity in the context of CiRM? The ACAP construct allows us assess whether or not organizations are able to effectively manage information flows. A high degree of potential ACAP helps public administrations understand what the citizens think, what they want, and what they might need. In turn, a high degree of realized ACAP allows public administrations to identify strategies and organizational means to optimize their operations with respect to an ever increasing quality of citizen relations.

4. RESULTS

While special phone services during emergency activations or hurricanes are not new to the State of Florida, 311 has broader implications on the County and its emergency and daily service operations.

311 improves accessibility to information and government services. Information is centralized and integrated seamlessly throughout the different channels and levels of government. 311 provides timely access to continuously updated information and services for the public and government entities. 311 data enables deeper insights into the public needs

and mind. This allows for a broader retrospective analysis when combined with data from other entities for continuous improvement. Miami-Dade government officials refer to the 311 system as an opportunity to continuously shape their government organization. Some look at 311, especially its data, as a key driver for organizational change, citizen orientation and multi-jurisdictional government. As we extract from our interviews, the 311 environment requires and trains a new type of administrative employee. 311 requires a system-level bureaucrat [3] that knows how to transform and exploit the knowledge that is available in the overall system.

5. CONCLUSION

With respect to literature on CiRM, we take the example of emergency management to illuminate the importance of effective knowledge and information management between government agencies and the broader public. By utilizing the absorptive capacity construct we have opened another avenue for future theoretical and empirical research on CiRM. For practitioners, we provided a real life case study on how efficient CiRM practices might look like. We furthermore tried to demonstrate that 311 is more than just a call center solution. As the case of Miami-Dade County shows, 311 is about adopting state-of-the art information management practices to build successful relationships between government agencies and its broader public in emergency and non-emergency situations.

6. ACKNOWLEDGEMENTS

We would like to thank all administrative members of Miami-Dade County under the leadership of George Burgess for their valuable feedback and access to 311 data. The researchers are partly sponsored by a Johann Wolfgang Goethe- University, Frankfurt am Main Doctoral Scholarship and the Rotary Foundation Ambassadorial Scholarship.

7. REFERENCES

[1]Cohen, Wesley M. / Levinthal, Daniel A. (1990). "Absorptive Capacity: A New Perspective on Learning and Innovation", *Administrative Science Quarterly*, 35, Special Issue: Technology, Organizations, and Innovation, 128-152.

[2]Mushkatel, Alvin H. / Weschler, Louis F. (1985). "Emergency Management and the Intergovernmental System", *Public Administration Review*, 45, Jan., 49-56.

[3]Reddick, Christopher G. (2005). "Citizen interaction with e-government: From the streets to the servers?" *Government Information Quarterly*, 22, 38-47.

What can e-Commerce and e-Government Learn from Each Other?

Hans J (Jochen) Scholl

The Information School, University of Washington

Box 352840, Seattle, WA 98195-2840, USA

1.206.685.9937

jscholl@u.washington.edu

Abstract

Comparative studies of the phenomena of e-Commerce/e-Business and e-Government have not yet emerged. However, such studies would most likely be instrumental in fostering cross-fertilization between the two evolutionary trajectories. They would further deepen the understanding of the sector-specific similarities and differences. Comparative research would also help establish performance measures and success criteria. This poster develops a vantage point for such a study.

Categories and Subject Descriptors

K.4 {Computers and Society}: K.4.2 *Organizational Impacts – K.4.4 Electronic Commerce*

D.2.11 {Software Engineering}: - Software Architectures - *Domain-specific architectures*.

General Terms

Design, Human Factors, Theory

Keywords

Comparative studies, e-Commerce, e-Business, e-Government, digital government, public-private distinction, traditional MIS and PMIS, metrics, performance measures, success factors, organizational impacts

1. Introduction

E-Commerce and e-Business have been portrayed as private-sector siblings and antecedents of e-Government (Scholl, 2003). At face value, many transactional and informational phenomena observed in private-sector e-Commerce and e-Business arenas seem to be mirrored in e-Government, and vice versa. Also, field-/front office work and back office work in both sectors have undergone major changes in the wake of introducing capabilities and methods, which exploit computer-mediated networks such as the Internet and the Web. Interestingly, few studies exist, if any, which capture the lessons learned and summarize current practices, or establish measures for success of this rapid socio technical evolution. Completely absent seem to be studies, which compare lessons learned, current practices, and established metrics across sector boundaries. With the

exception of technological advances in systems and networking, e-Business and e-Commerce developments in one sector remain mostly unattended in the other sector; and few, if any, spillover effects seem to exist. The organizational and social impacts and consequences in the e-Commerce/e-Government evolution, however, may represent the far more challenging problems to understand than the admittedly non-trivial (technical) interplay of systems and networks. This poster suggests that comparative research on the differences and similarities of e-Commerce and e-Government, the respective lessons learned, the current practices, and the metrics of success would benefit both sectors and both evolutionary paths. Deeper understanding of the complex mesh of technological, organizational, and social factors and processes in both e-Commerce and e-Government might lead to practice-relevant cross-fertilization and reduction of unnecessary reduplication.

2. The Public-Private Distinction

Traditionally, public-to-private differences have been identified in three areas: (1) environmental drivers and constraints, (2) organizational mandates and scope, and (3) internal processes, complexities, and incentives (Rainey *et al.*, 1976). Among other distinguishing dimensions, the private sector has been praised for its higher agility, greater resourcefulness, less burdensome bureaucracy, and stronger motivation to proactively innovate when compared with public sector organizations (Boyne, 2002; Bozeman & Bretschneider, 1986; Bretschneider, 1990; Mohan & Holstein, 1990; Rainey *et al.*, 1976). However, as Perry and Rainey observed, in practice the distinction between the two sectors might be troublesome, since apart from the clear-cut cases of publicly owned, and publicly funded organizations versus privately owned, and privately funded firms, all shades of grey, that is, hybrids exist, which considerably blur the sector boundaries and make the distinction between public and private sector organizations less meaningful, if not even problematic at times (Perry & Rainey, 1988). When studying the similarities and differences between e-Commerce and e-Business, on the one hand, and e-Government on the other hand, the wide spectrum of hybrid formats needs to be taken into account. It might be informative to analyze how the differences and similarities change along the public-to-private continuum.

3. Sector Differences and Similarities as Seen by Traditional Information Systems (IS) Research

In contrast to private Management Information Systems (MIS), which are mainly geared to increasing economic efficiency and profitability, their public counterparts (PMIS) had to

simultaneously provide both economic and political efficiencies and also serve a policy mission (Bozeman & Bretschneider, 1986). Further, unlike private MIS, PMIS yielded only few labor savings. Finally, the (at times extreme) scarcity of skilled labor had been a chronic challenge for public IT managers and CIOs for decades (*ibid*).

To a certain extent, those differences might also hold when comparing e-Commerce/e-Business with e-Government, and vice versa. In a recent study, public IS managers were found to pursue different priorities in their IT investments than their private colleagues: In rank order, the top five priorities for public-sector CIOs were (1) the implementation of an IT architecture, (2) cultural change, (3) hiring/retaining skilled professionals, (4) unifying/streamlining, and (5) simplifying business processes (Ward & Mitchell, 2004). By contrast, private-sector CIOs ranked their priorities as (1) simplifying business processes, (2) improving service, (3) effective relationships with senior executives, (4) preventing intrusions, and (5) the implementation of an IT architecture (*ibid*).

Traditional PMIS and e-Government systems, although subjected to probably identical sector specifics, appear to be markedly different in a number of ways (Scholl, 2005b). Therefore, a detailed analysis of the characteristics of e-Government systems, applications, and of their organizational impacts compared with private-sector e-Commerce systems, applications, and their respective organizational impacts appears necessary for developing deeper understanding.

4. Known Impacts, Effects, and Proposed Measures of Success

The e-Commerce/e-Business and the e-Government-related literatures have independently reported on effects and outcomes, which indicate similarities, differences, as well as different emphases in the two sectors. Increased trust, risk and knowledge sharing, improved economies of scale and scope, shortened time to market, lower transaction cost, reduced information asymmetries, and improved coordination were found among the effects in the private sector as a consequence of the introduction of e-Commerce (Amit & Zott, 2001; Porter, 2001). E-Government has reportedly led to business process change, process speedups, increased internal efficiency, improved information sharing and interoperation, greater transparency and accountability, greater proximity to citizens, improved service levels among other effects (Guizarro, 2004; Kaylor, 2005; Scholl, 2003, 2005a; Traunmüller & Wimmer, 2002). The literature on measuring the performance (and the success) of e-Commerce/e-Business and e-Government, however, is still in its early stages of development. Some scholars propose the extension of existing frameworks (DeLone & McLean, 2004) for measuring the relative performance, while others propose holistic frameworks based on an augmented balanced scorecard approach, which capture the organizational and strategic dimensions (Bremser & Chung, 2005).

5. Conclusion

Comparative research on e-Commerce/e-Business, on the one hand, and e-Government, on the other hand will improve the understanding of (a) current practices and lessons learned in the two sectors, (b) the potential for cross-fertilization between the sectors, (c) the nature and origins of both similarities and differences between the evolutionary trajectories of the two phenomena, and (d) the relative effectiveness and performance of e-Commerce and e-Government. A future proposal will

meticulously detail an empirical research design, which covers the research direction charted out above.

6. References

- [1] Amit, R., & Zott, C. (2001). Value creation in e-Business. *Strategic Management Journal*, 22(6-7), 493-520.
- [2] Boyne, G. A. (2002). Public and private management: What's the difference? *Journal of Management Studies*, 39(1), 97-122.
- [3] Bozeman, B., & Bretschneider, S. (1986). Public management information systems: Theory and prescriptions. *Public Administration Review*, 46(November (special issue)), 475-489.
- [4] Bremser, W. G., & Chung, Q. B. (2005). A framework for performance measurement in the e-business environment. *Electronic Commerce Research and Applications*, 4(4), 395-412.
- [5] Bretschneider, S. (1990). Management information systems in public and private organization: An empirical test. *Public Administration Review*, 50(September/October), 536-545.
- [6] DeLone, W. H., & McLean, E. R. (2004). Measuring e-Commerce success: Applying the DeLone & McLean information systems success model. *International Journal of Electronic Commerce*, 9(1), 31-47.
- [7] Guizarro, L. (2004). Analysis of the interoperability frameworks in e-Government initiatives. In R. Traunmüller (Ed.), *Electronic government: Third international conference, egov 2004, Zaragoza, Spain, August 30-September 3, 2004. Proceedings* (Vol. 3183, Lecture notes in computer science, pp. 36-39). New York, NY: Springer-Verlag Berlin Heidelberg.
- [8] Kaylor, C. H. (2005). The next wave of e-Government: The challenges of data architecture. *Bulletin of the American Society for Information Science and Technology*, 31(2), 18-22.
- [9] Mohan, L., & Holstein, W. K. (1990). Eis: It can work in the public sector. *MIS Quarterly*, 14(4), 434-448.
- [10] Perry, J. L., & Rainey, H. G. (1988). The public-private distinction in organization theory: A critique and research strategy. *Academy of Management Review*, 13(2), 182-201.
- [11] Porter, M. E. (2001). Strategy and the Internet. *Harvard Business Review*, 79(3), 63-78.
- [12] Rainey, H., Backoff, R., & Levine, C. (1976). Comparing public and private organizations. *Public Administration Review*, 36(2), 233-244.
- [13] Scholl, H. J. (2003, 1/6 to 1/10). *E-Government: A special case of business process change*. Paper presented at the 36th Hawaiian International Conference on System Sciences (HICSS36), Waikoloa, HI.
- [14] Scholl, H. J. (2005a). E-government-induced business process change (BPC): An empirical study of current practices. *International Journal of Electronic Government Research*, 1(2), 25-47.
- [15] Scholl, H. J. (2005b). Electronic government: Information management capacity, organizational capabilities, and the sourcing mix. *Government Information Quarterly*, Forthcoming.
- [16] Traunmüller, R., & Wimmer, M. A. (2002, Oct 29-31). *Integration: The next challenge in e-Government*. Paper presented at the EurAsia-ICT 2002, Shiraz, Iran.
- [17] Ward, M., & Mitchell, S. (2004). A comparison of the strategic priorities of public and private sector information resource executives. *Government Information Quarterly*, 21(3), 284-304.

Virtualization Technologies in Transnational DG

Maurício Tsugawa, Andréa Matsunaga and José A. B. Fortes
ACIS Laboratory, Dpt. of Electrical and Computer Eng., University of Florida
PO Box 116200, Gainesville, FL, 32611-6200, USA
1 (352) 392-4964

{ tsugawa, ammatsun, fortess}@acis.ufl.edu

I. Introduction

Naïve deployments of digital government (DG) systems across organizations in different countries inevitably face severe technical, sociopolitical and economical barriers. Some of these barriers are the result of independently created IT infrastructures with distinct use-policies, varying functional capabilities and different interoperability requirements. In general, IT heterogeneity is inevitable as it results from differences in economical and technical capabilities of the countries, differences in agency missions, distinct regulatory contexts (which may, for example, specify what kind of software must be used) and unequal human IT resources. Deployment of an available DG system into an existing infrastructure may require use of new and/or existing hardware and/or software at different locations, processing and accessing data located in distinct agencies, and communication among many IT entities. In this context, heterogeneity can lead to several forms of incompatibilities, namely hardware, software, communication, data, security and accessibility. A small subset of aspects of IT infrastructures where there might be important differences includes the following: operating systems, hardware, firewall mechanisms and policies, software applications and authorization procedures.

Deployment of a transnational DG system either requires extensive negotiation and agreement among the participants to change their IT infrastructure and policies as needed, or a solution that hides away the heterogeneity of the IT environments. That social science or political skills can by themselves succeed along the lines of the first approach is wishful thinking. This paper argues for the necessity and viability of the latter solution on the basis of existing and new virtualization technologies introduced by the authors and experiential evidence in the context of a transnational DG project[1].

We argue that virtualization technologies can be used to greatly decouple the deployment of trans-organization DG from organization IT architecture and policies. They do so by creating IT environments that are homogeneous across institutions and specifically suited to the DG applications of interest. For example, in the context of a transnational DG (TDG) project [1], these applications included database query systems, machine translation

software and conversational interfaces and the challenge was to deploy them across three countries (Belize, US and the Dominican Republic) with vastly different IT infrastructures.

The key outcome of using virtualization technologies is the ability to overlay a homogeneous (virtual) infrastructure across heterogeneous (physical) infrastructures. This requires technologies for virtualizing computers, networks and applications. Virtual computers, more commonly called virtual machines (VMs), can currently be implemented using a variety of commercial or publicly available technologies. To the best of our knowledge, we are the first to propose and describe the contexts and benefits of their use in transnational DG in Section II (see also [2]). Section III refers to our unique approach to the implementation of virtual networks as needed for interconnecting both VMs and physical machines across different countries. Finally, Section IV considers our solution for virtualizing applications as needed to reuse existing applications in different DG contexts.

II. Virtual Machines (VM)

Machine virtualization, as recently revisited in [3][4], enables a single physical machine to simultaneously behave as multiple isolated independent (virtual) machines on which distinct operating systems and software can be deployed. Several aspects of VMs make them essential for DG applications:

- **Compatibility:** DG system software components whose execution environments are incompatible with existing IT infrastructure can be deployed on VMs where the execution environment is recreated.
- **Interoperability:** When integrating applications into a DG system, conflicting software library requirements can make it impossible to run all needed components in a single platform. The integration process is facilitated by the possibility of assigning independent VM platforms to run the components.
- **Security:** VMs exhibit a high level of isolation from each other. A compromised application is highly unlikely to affect other components of a system, given that they run in different VMs. Even if an application causes its assigned VM to crash, other VMs and applications will not be affected.
- **Testing and debugging:** VMs allow changes in hardware parameters by editing a single configuration file. Prior to deployment, changes in memory, processor and storage configuration can be tried in order to catch software problems that might only affect certain hardware configurations.
- **Version control:** System-level versioning can be accomplished by taking advantage of VM's checkpointing and cloning capabilities. If a system update fails, it is possible to go back to the previous working condition very efficiently.

- **Rapid deployment:** There is no need to modify software components so they can run in the IT infrastructures. VMs can recreate the necessary development environments and encapsulate all the needed software components
- **Replication:** VMs can be cloned and deployed in a way similar to file copying. Systems can be replicated very easily, and training of human resources is facilitated by the ability to make fully operational systems available to each individual or government agency.
- **Independence from local IT infrastructure operations:** Physical infrastructure changes, resulting, for example, from maintenance or upgrades, do not affect the software running on VMs.
- **Low cost:** Hardware support for virtualization [5] and free VM software [6][7] make it possible for VMs to be deployed without additional cost.

III. Virtual Networks (VNs)

VNs allow the creation of independent and isolated networks across the public Internet infrastructure and private networks. ViNe [8] is an example of a virtual networking technique that has been developed by the authors. It enables any-to-any communication among physical and virtual machines, even if the connectivity of the physical networks is limited by the presence of firewalls. The benefits of VNs are similar to those of VMs and, in addition, include also the following

- **Any-to-any communication:** DG systems that integrate software components are distributed in nature and, in general, are designed assuming symmetric connectivity. This symmetric connectivity is not available when organizations limit access to their internal networks through firewalls or NATs. ViNe securely enables firewall and/or NAT traversals transparently thus providing an application-friendly network environment that meets the symmetry requirement.
- **Low administration overhead:** connecting networks in different administrative domains involves work, negotiation and information exchange by administrators from all domains. Some organizations do not have means to execute some network operations due to limitations imposed by network providers (e.g., lack of connectivity to the public network and/or static address assignment). ViNe is designed to be largely self-organized/managed requiring low administrative interventions once deployed.
- **Independence from shared network limitations:** the virtualization indirection simplifies networking issues - such as quality-of-service, reliability and security - that are hard to address in shared physical infrastructures.
- **Platform independence:** ViNe is designed to work with any platform without the need for additional software.

IV. Virtual Applications (VAs)

VAs allow the integration of isolated applications (applications that are not ready to interoperate with other applications through the network) by creating a homogeneous application communication layer based on standards like Web-services. In addition, as exemplified by middleware developed by the authors to create Virtual Applications Services [9], VAs benefits include also the following:

- **Rapid generation and deployment of VAs:** The process of generating, building and deploying an application with command-line interface as a service is fully automated (starting from the syntax of the command line).
- **Customization:** An application interface can be customized to appear differently from the original application, allowing several versions to exist for different purposes.
- **Validation:** Validation of service inputs can be done thus improving the usability of the service and preventing the propagation of errors.
- **Composition:** A VA can consist of many VAs, thus naturally supporting the creation of new applications from existing ones.

V. Conclusion

This article summarizes the rationale and benefits of using virtualization technologies in enabling the dynamic deployment of virtual IT infrastructures across countries with arbitrary dissimilar infrastructures. The resulting virtual IT environments can be created to suit the needs of a specific DG system software. Public and commercial offerings of VM technologies, along with emerging hardware support in mainstream architectures, will turn VMs into pervasive capabilities of computers in the near future. They will be complemented by technologies currently being researched for virtual networks and virtual applications. Existing prototypes developed by the authors in the context of a TDG project confirm both the viability and the advantages of such technologies. We thus believe that the use of virtualization should be further researched and fully leveraged to address the technical challenges of efficient and dynamic deployment of TDG systems.

Acknowledgements: This research is funded in part by NSF awards EIA-0107686 and EIA-0131886. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References:

- [1] Su, S., et. al., Transnational Information Sharing, Event Notification, Rule Enforcement and Process Coordination. International Journal of Electronic Government Research (IJEGR), Vol. 1, No. 2, April-June, 2005.
- [2] Matsunaga, A., et. al., The Case for Virtual Machines in Transnational Digital Government, TR-ACIS-05-009, 2005.
- [3] Figueiredo, R., et. Al., Resource Virtualization Renaissance. Computer 38, 5, May 2005, p.28-31.
- [4] Smith, J. E. and Nair, R. Virtual Machine Versatile Platforms for Systems and Processes. Morgan Kaufmann Publishers, June 2005, 638p.
- [5] Intel. Intel Vanderpool Technology for IA-32 Processors (VT-x) Preliminary Specification. Intel, January, 2005.
- [6] VMware, Inc. Introducing VMware Virtual Platform. Technical white paper, February, 1999.
- [7] Barham, P., et. al., Xen and the Art of Virtualization. In Proceedings of SOSP 2003, October, 2003.
- [8] Tsugawa, M., and Fortes, J.A.B., A Virtual Network (ViNe) Architecture for Grid Computing. In Proceedings of 20th IPDPS, Greece, April, 2006. (to appear).
- [9] Matsunaga, A., et. al., Science gateways made easy: the In-VIGO approach. Concurrency and Computation: Practice and Experience, Wiley Press, 2006. (in press).

An Electronic Social Network to Market Topics of Public Interest: Net@INA

L. Valadares Tavares

President of INA; Full Professor of Systems and Management of the Technical University of Lisbon (IST).
Palácio dos Marqueses de Pombal
2784-540, Oeiras, Portugal
00-351-214465420/19

lvt@ina.pt

Paulo Silva

Coordinator of Marketing for INA; MBA at the New University of Lisbon
Palácio dos Marqueses de Pombal
2784-540, Oeiras, Portugal
00-351-214465387

Paulo.silva@ina.pt

ABSTRACT

In this paper, the authors present a model to develop an electronic network devoted to market contents of public interest (Topics of Public Interest) based on the experimental electronic network already in action at INA (WWW.INA.PT).

The proposed model can be the beginning of a new front of developments within E-Government with the purpose of improving general information and knowledge about key topics to support the implementation of public policies requiring difficult processes of change.

Categories and Subject Descriptors

C.2.3 [Network Operations]: Public networks; D.4.8 [Performance]: Modeling and prediction.

General Terms

Management, Measurement, Documentation, Design, Experimentation, Human Factors.

Keywords

Social Networks; E-Government; Topics of Public Interest.

1. INTRODUCTION

Nowadays, societies are facing new challenges to cope with the processes of globalization, increasing demand for quality of life, sustainability of development, stable employment, technological innovation, environmental equilibrium and migration dynamics.

These challenges are tackled by a wide spectrum of public policies aiming to drive the required processes of societal change. Such policies are converging to a strong political consensus embracing different streams of political thought and emphasizing the need:

- a) to reconsider the role of the State avoiding the production of goods or services and focusing on the functions of planning, contracting, controlling, regulating and evaluating.
- b) to give priority to the development of human capital and innovation adopting the new paradigms of knowledge society.
- c) to contribute to stable processes of development respecting environmental equilibrium and improving the quality of life.

- d) to promote social cohesion through the development of social markets where public resources are used to acquire social services provided by firms, agencies or not-for profit organizations (churches, societies, etc.).

These priorities imply a new style of communication and knowledge share on topics of public interest involving government, public administration and civil society in order that the required changes will be understood and supported.

Actually, the existing society has two major instruments to implement such processes:

- a) the educational system, which is supposed to be devoted to the promotion of values and of knowledge on topics of public interest.
- b) media (newspapers, magazines, radio, TV) exposing everyday population to multiple streams of news and opinions.

Education tends to be less updated regarding new challenges and media are dominated by the war of audiences leading their contents to more futile matters.

This explains the need to develop alternative systems of communicating with the purpose of improving general understanding about public needs and justifying the political processes of change.

Therefore the question studied on this paper is: can the new digital technology contribute to this purpose?

How much contribution is being offered by E-Government?

Which other initiatives can be developed?

2. TOPICS OF PUBLIC INTEREST

The taxonomy of knowledge is a key concept to pursue the paradigms of knowledge society and the concept of Topic of Public Interest (TPI) can play a key role on public policies:

TPI can be defined as a set of organized data, reports, evaluation criteria, comparative analyses, opinions and comments, discussions and proposals about an issue of public interest deserving the attention of policy makers to conceive, to design and to implement a public policy to improve the existing situation.

These topics require a systematic and professional treatment to collect and to analyse the appropriate material and to stimulate discussion fora and consistent proposals.

The considered issues cover any subject connected to problems or aspirations including a societal dimension and requiring public intervention such as improvement of local security, reduction of pollution levels, renewal of historical urban centres, integration of multi-mode transportation systems, access modes for health care networks, social support immigrants, etc.

Better information and knowledge about such topics can be critical to improve policy design and to increase public support to apply a new public policy avoiding the dominance of those loosing advantages from the implementation of such policy.

3. NET@INA

This network was started on February 2003 to share information, knowledge and on-line services provided by INA to any person or organization interested on public management.

Also, a monthly newsletter is provided free of charge to any member.

More than 4000 members interact through this network on topics of public interest giving top priority to the following contents:

- a) Courses and seminars 48%
- b) Information about public decision 42%
- c) Events (news and comments) 13%
- d) Editions 7%

Almost half of the members have no more than 35 years of age (46%) and about 40% between 35 and 55.

This networking is showing a high level of vitality and therefore the new model of social electronic network will be the next step of evolution of this system.

4. THE PROPOSED MODEL FOR AN ELECTRONIC SOCIAL NETWORK

E-Government has been rapidly expanding to serve two major purposes:

- a) presentation of documents and data relevant for the relationship between Government and citizens or firms (steps and forms to apply for a specific licence, conditions and dates to receive proposals concerning a public procurement process, land use restrictions, etc.).
- b) implementation of interactive administrative procedures.

This is the case of specific cases such as the application for some bureaucratic documents (identity cards, driving licences, etc.).

The authors believe that the promotion of information and knowledge about Topics of Public Interest can be effectively pursued using the electronic instruments and therefore a model is proposed to achieve this objective.

This model (Figure 1) is based on the electronic network already existing at INA (National Institute of Public Administration of Portugal) and includes the following components:

A) Specialized Sources

These sources can include a wide variety of social actors such as experts, officers, opinion leaders, journalists.

The sources provide their contribution through the electronic network.

B) Accreditation units

This model implies units devoted to check the credibility and reliability of contents received from sources.

C) Analysts

These actors provide analytical frameworks and evaluation criteria to map and to compare the provided contents.

D) Knowledge Providers

These units are supposed to use the output of C) to produce knowledge contents to be shared by the electronic network.

These products will include different classes of knowledge:

- Facts;
- Problems;
- Issues;
- Alternatives;
- Criteria;
- Comparisons;
- Debates;
- Options.

5. FINAL REMARKS

The existing Net@INA will be gradually expanded to implement the model presented in Figure 1 and experimental results will be provided in a future publication.

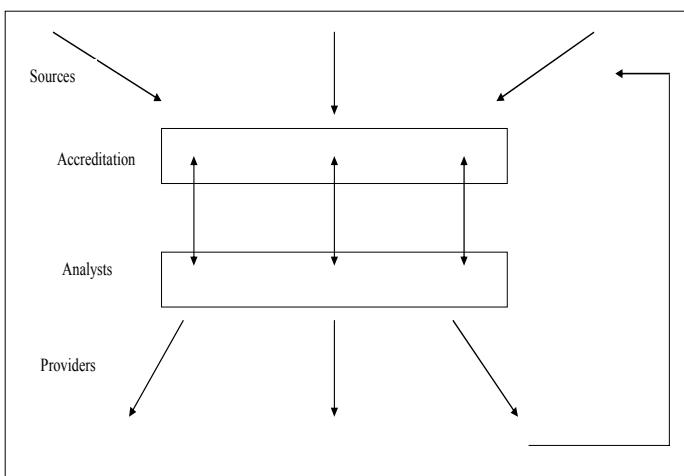


Figure 1. A Model of Social Electronic Network

A Performance Ratings Framework for the Evaluation of Electronic Voting Systems

Project Highlights: Exploratory Research

Poorvi L. Vora (PI)

Dept. of Computer Science
George Washington University
Washington, DC, USA
poorvi@gwu.edu

Rahul Simha

Dept. of Computer Science
George Washington University
Washington, DC, USA
simha@gwu.edu

Jonathan Stanton

Dept. of Computer Science
George Washington University
Washington, DC, USA
jstanton@gwu.edu

ABSTRACT

This paper outlines the highlights of the DG-funded year-long project on performance-ratings for voting system evaluation. It describes recent work on evaluating voting systems with respect to privacy, and on a performance evaluation of an implemented system. It ends with a summary of plans for the next few months.

1. INTRODUCTION

Voting is arguably one of the most important community activities in democracies, and the entire voting process could benefit from advances in information technology. However, in spite of the fact that cryptography has enabled dramatic changes in electronic commerce and banking, cryptographic voting schemes have made almost no impact on the voting technology available to the voter. A large part of the reason is elucidated in [3]: the current standard for electronic voting is pass/fail in nature, is based on criteria defined for earlier-generation electromechanical devices, and therefore provides no incentive for the use of modern cryptographic techniques. A more useful standard would be one based on a set of desirable voting system properties, and a means of rating the performance of a system with respect to each property. The goal of this project is to carry out the research underlying the development of such a performance-rating system.

This paper describes several approaches to analyzing voting systems with respect to the desirable qualities of voter privacy and system efficiency. We outline briefly an information-theoretic privacy measure that we have described in detail in our DG-2005 paper [1]. We also summarize the material in our paper [2] under review, where we present Citizen-Verified Voting (CVV), our implementation of a Chaum-voting system, describe design decisions we took to avoid the insertion of unintended covert channels, and evaluate the vote processing efficiency of the implementation. Further, this paper also describes plans to make available an Open Source version of the CVV prototype, and to hold a workshop on performance ratings of voting systems.

2. RESEARCH

2.1 A Privacy Measure

With Lillie Coney (EPIC), Joe Hall (UC Berkeley) and David Wagner (UC Berkeley) we have provided a definition of perfect privacy of a voting system, and a measure to quantify privacy loss when the system does not provide perfect privacy. The measure is entropy-based, and can be used to measure entire voting systems and processes, whether paper-based or electronic. For example, it can be used to measure the privacy of a given voting system with and without voter collusion, or it can be used to measure the difference in the privacy provided when entire cast ballots are released as against that provided when only vote counts are released.

2.2 Covert Channel Analysis

The existence of covert channels that convey information about an individual's vote is a significant concern in voting systems, which have strong privacy requirements. While the core voting protocols provide provable privacy to voter's choices under certain assumptions, there is typically still the potential for specific implementations to allow covert channels originating at the polling machine. The polling machine sees unencrypted votes, and can often link them to some information identifying the voter – a serial number, for example, or the order in which the vote was received, or the time of the vote. It could thus be the source of leakage of information on a voter's vote. Before we describe potential covert channels in CVV and design decisions to avoid them, we briefly describe CVV.

2.2.1 CVV: *Citizen-Verified Voting*

CVV is an implementation of a voting system proposed by David Chaum. Using this system, a voter can convince herself that her vote was correctly recorded and counted, and a polling machine that cheats to change the vote count can be caught with high probability. CVV uses an encrypted receipt provided to the voter. The receipt allows the voter to check that her vote is counted correctly, yet does not allow her to prove how she voted. The voter can check that the receipt is in the collection of encrypted receipts that will be counted; these would typically be made available on a website. The entire set of receipts can be checked to ensure each was correctly constructed. The collection of encrypted receipts goes serially through a group of trustees; each trustee partially decrypts each receipt and shuffles the entire set.

The trustees thus, together, provide a set of decrypted receipts which may be counted by anyone. If at least one trustee performs an honest shuffle, a specific vote cannot be traced back to the voter. The trustees are later audited, and trustees who do not perform the decryptions correctly are caught with high probability.

2.2.2 Covert Channels in CVV

The two main covert channel risks in CVV are (a) random values that are used in the protocol, as a party can make a non-random choice – that appears random – in order to communicate some information, and (b) a semantic channel in which visually unnoticeable information is added to the image of the voter’s choices by the polling machine. In the semantic channel, the covert information can be read by anyone after the votes are fully decrypted and posted in clear text for counting. In this section we discuss specific approaches to avoiding random values in the official messages. For the semantic channel, we have specified a specific font and spatial positioning of characters and data that allows the visual information to be validated as not containing any non-specified information.

Random numbers are used in the protocol in several places. For example, random numbers are used (a) when the polling machine encrypts the vote to provide the voter a receipt, and (b) when it adds, to the encrypted vote, values used by the trustees to decrypt the vote. In a typical implementation of CVV, each value is encrypted with a symmetric (private/secret) key, which is itself encrypted with the public key of the corresponding trustee. The secret key is randomly generated, and can constitute a covert channel between the polling machine and the corresponding trustee. Note that this is an issue that also plagues public key encryption in general, and not just the CVV algorithm. Further, the initialization vector, used in every symmetric encryption, can be used as a covert channel in the same manner, as can be any other encryption parameter that can be randomly chosen. To detect the use of these two covert channels, and to thus discourage their use, we propose that the parameters used for encryption – the symmetric key and the initialization vector, for example – be generated in a deterministic manner from the serial number of the vote. Another source of randomness is in the trustee’s choice of a permutation. Our approach is to have the trustee commit to the permutation before the voting is performed, so that the permutation reveals no information on individual votes.

2.3 Tools

We have developed software tools that can automate the testing of electronic voting systems for both efficiency and correctness. These tools are only preliminary, but can generate a large set of standard ballots, run them through a software-based voting system, and verify the correctness of the election outcome. They have also produced an initial performance evaluation of the CVV implementation. For example, these tools demonstrate that the implementation takes about one hundred seconds to audit one thousand votes (see Figure 1).

3. FUTURE PLANS

We aim to use the privacy metric of Section 2.1 as the basis of a Privacy Measurement Criterion (PMC) that we will

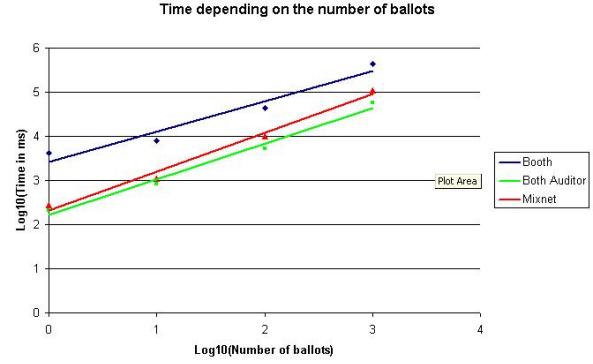


Figure 1: Efficiency Plot

propose to the PMC Working Group of the Voting Systems Performance Rating (VSPR) organization of which one of the PIs is Chair. We also plan to have a small workshop on the development of performance ratings criteria. Talks are currently ongoing with NIST regarding whether they would wish to be involved; NIST organized a Threat Analysis workshop last year, the output of which would be very useful for performance ratings efforts. Further, performance ratings, or evaluations of voting systems with respect to specific threats, would appear to be a natural next step to a Threat Analysis effort. Lastly, we will make available an open-source version of CVV.

4. CONCLUSIONS

The project will help refine the debate on voting systems by defining a technical core for ratings of performance with respect to specific properties. Further, the open-source release of a robust voting system with strong voter protection properties would meet a wide variety of voter requirements, and therefore has the potential of attracting other research efforts, raising awareness and, over time, helping increase public trust in electronic voting systems. We hope that, through the workshop and the interaction with NIST, the project will contribute to the process of determining a new voting standard. The project is also helping train graduate students in the area of security, and, through VSPR, has resulted in the development of ties between the computer science community and public interest groups engaged in democratic processes.

5. REFERENCES

- [1] L. Coney, J. L. Hall, P. L. Vora, and D. Wagner. Towards a privacy measurement criterion for voting systems. In *National Conference on Digital Government Research*, May 2005.
- [2] B. Hosp, S. Popoveniuc, R. Simha, J. Stanton, and P. Vora. Implementation and evaluation of a cryptographically secure voting system. In Review, December 2005.
- [3] P. L. Vora, B. Adida, R. Bucholz, D. Chaum, D. L. Dill, D. Jefferson, D. W. Jones, W. Lattin, A. D. Rubin, M. I. Shamos, and M. Yung. Evaluation of voting systems. *Comm. of the ACM*, Nov. 2004.

A Probabilistic Model for Approximate Identity Matching

G. Alan Wang
University of Arizona
1130 E. Helen St., Rm 430
Tucson, AZ 85721
+1 (520) 621-2748
gang@eller.arizona.edu

gang@eller.arizona.edu

Hsinchun Chen
University of Arizona
1130 E. Helen St., Rm 430
Tucson, AZ 85721
+1 (520) 621-2748

hchen@eller.arizona.edu

Homa Atabakhsh
University of Arizona
1130 E. Helen St., Rm 430
Tucson, AZ 85721
+1 (520) 621-2748

homa@eller.arizona.edu

ABSTRACT

Identity management is critical to various governmental practices ranging from providing citizens services to enforcing homeland security. The task of searching for a specific identity is difficult because multiple identity representations may exist due to issues related to unintentional errors and intentional deception. We propose a probabilistic Naïve Bayes model that improves existing identity matching techniques in terms of effectiveness. Experiments show that our proposed model performs significantly better than the exact-match based technique as well as the approximate-match based record comparison algorithm. In addition, our model greatly reduces the efforts of manually labeling training instances by employing a semi-supervised learning approach. This training method outperforms both fully supervised and unsupervised learning. With a training dataset that only contains 10% labeled instances, our model achieves a performance comparable to that of a fully supervised learning.

Categories and Subject Descriptors

G.3 [Probability and Statistics]: Probabilistic algorithms

General Terms

Algorithms

Keywords

Identity matching, Naïve Bayes model, semi-supervised learning

1. INTRODUCTION

Many governmental agencies manage identity information for various purposes ranging from providing citizens services to enforcing homeland security. Identity verification is a common practice and is used to verify whether a person is who he/she claims to be. This is, however, a surprisingly complex problem [1]. In this research we intend to carefully examine what causes uncertainty in matching identities and propose a probability-based Naïve-Bayes model for approximate identity matching.

2. LITERATURE REVIEW

2.1 Identity Problems

In a previous case study we studied real identity records associated with 200 criminals in the Tucson Police Department (TPD) [7]. We found that errors and deception result in multiple identity representations for an individual person in one system or across multiple systems. We created a taxonomy of the identified problems (Figure 1). Among other identity features, name, date-of-birth (DOB), ID numbers (e.g., social security numbers), and address were found to indicate deception or errors in most cases.

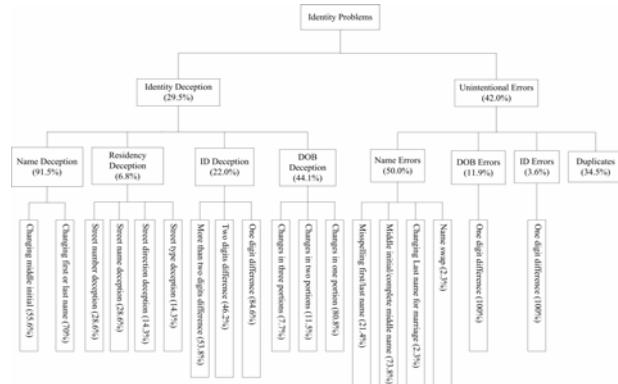


Figure 1. Taxonomy of identity problems.

2.2 Identity Matching Techniques

To the best of our knowledge, there are few solutions proposed for the problem of identity matching. Marshall et al. [3] provided an exact-value matching technique for law enforcement applications. Two identities are considered matching only if their name and DOB values are identical. However, as identity information is unreliable, it is possible that identities referring to the same person have disagreeing values. Wang et al. [6] proposed a record comparison algorithm to detect deceptive identities. Two identities being compared are considered matching when an overall similarity rating is greater than a threshold value. A supervised training process is required to determine the threshold value. However, supervised training requires experts to manually generate a training dataset. This manual process can be time-consuming and inefficient.

2.3 General Entity Matching Techniques

The problem of identity matching can be considered as a special case of entity matching where it is determined whether an entity in one database is the same as the entity in another one [2].

Ravikumar and Cohen proposed a three-layer hierarchical graphical model for entity matching [5]. This approach employed unsupervised learning to avoid human intervention. However, unsupervised learning has been shown to be insufficient for training [4] and the model is subject to over-fitting the noisy data [5].

3. RESEARCH DESIGN

3.1 A Multi-Layer Naïve Bayes Model

Based on the three-layer hierarchical graphical model, we propose a multi-layer Naïve Bayes model by removing the dependencies between latent variables and allowing the capture of complex matching heuristics. We argue that semi-supervised learning will achieve a balance between training accuracy and human intervention. We propose a multi-layer Naïve Bayes model for identity matching as shown in Figure 2. According to our case study on identity problems, matching values in name, DOB, ID numbers, and address indicate matching decisions in most cases. We have chosen these four features to represent each identity in our proposed model.

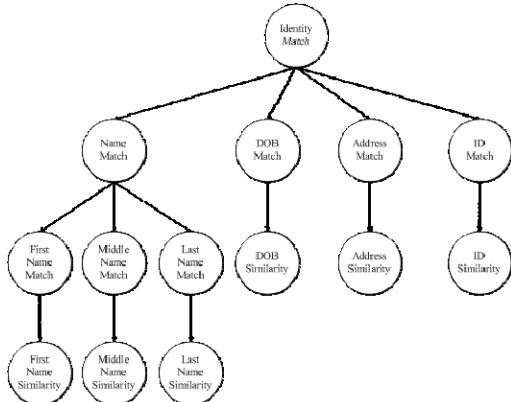


Figure 2. A probabilistic model for identity matching.

4. EXPERIMENTS

We used real identity records stored in the TPD database as our testbed. We asked a police veteran to manually verify the possible matching records of 200 suspects. The training dataset was generated by comparing each suspect's primary identity against each of its possible matches. The training dataset contained 20,000 comparisons, each of which was represented by a comparison vector and a matching decision label. We measured the performance of our proposed model using the following measures: recall, precision, and F-measure. Those measures are widely used in information retrieval.

4.1 Experimental Design

In our experiments we compared the performance of our proposed model against that of other existing identity matching techniques, namely the exact-match based technique [3] and the record comparison algorithm [6]. We also evaluated the performance differences of our model in three different learning modes: supervised, semi-supervised, and unsupervised learning.

4.2 Experimental Results

Our proposed model using either supervised, semi-supervised, or unsupervised learning, performed significantly better than the

exact-match technique ($p\text{-values} < 0.001$) and the record comparison algorithm ($p\text{-value} < 0.05$). We also found that the F-measure of semi-supervised learning remained high (0.8580) when only 10% of training instances were labeled. This could greatly reduce human interventions while maintaining an effective matching performance. Unsupervised learning significantly underperformed semi-supervised learning at all levels ($p\text{-values} < 0.001$).

5. CONCLUSIONS

Identity information is critical to various organizational practices ranging from customer relationship management to crime investigation. We proposed a probabilistic Naïve Bayes model that improved existing identity matching techniques in terms of effectiveness. We also achieved a balance between training effectiveness and human intervention. In the future we intend to develop an identity matching tool that supports various decision making processes involving identity information. Such a system is expected to improve the ability of information processing and sharing in various governmental agencies.

6. ACKNOWLEDGMENTS

This project has primarily been funded by the following grant:

Department of Homeland Security (DHS), "BorderSafe: Cross Jurisdictional Information Sharing, Analysis, and Visualization," September 2003-August 2005.

7. REFERENCES

- [1] Camp, J., "Identity in Digital Government," presented at 2003 Civic Scenario Workshop: An Event of the Kennedy School of Government, (Cambridge, MA 02138, 2003)
- [2] Dey, D., Sarkar, S., and De, P., "A Distance-Based Approach to Entity Reconciliation in Heterogeneous Databases," *IEEE Transactions on Knowledge and Data Engineering*, vol. 14 2002, pp. 567-582, 2002.
- [3] Marshall, B., Kaza, S., Xu, J., Atabakhsh, H., Petersen, T., Violette, C., and Chen, H., "Cross-Jurisdictional criminal activity networks to support border and transportation security," presented at 7th Annual IEEE Conference on Intelligent Transportation Systems (ITSC 2004), (Washington, D.C., 2004)
- [4] Nigam, K., McCallum, A. K., Thrun, S., and Mitchell, T., "Text Classification from Labeled and Unlabeled Documents using EM," *Machine Learning*, vol. 39 2000), pp. 103-134, 2000.
- [5] Ravikumar, P. and Cohen, W. W., "A Hierarchical Graphical Model for Record Linkage," presented at 20th Conference on Uncertainty in Artificial Intelligence (UAI '04), (Banff Park Lodge, Banff, Canada, 2004)
- [6] Wang, G., Chen, H., and Atabakhsh, H., "Automatically Detecting Deceptive Criminal Identities," *Communications of the ACM*, vol. 47 2004, pp. 71-76, 2004.
- [7] Wang, G. A., Atabakhsh, H., Petersen, T., and Chen, H., "Discovering Identity Problems: A Case Study," in *Intelligence and Security Informatics: IEEE International Conference on Intelligence and Security Informatics (ISI 2005)*. Atlanta, GA, 2005.

Semantic Web Technologies to Automate Searching for Geospatial Data

Nancy Wiegand

University of Wisconsin-Madison

550 Babcock Drive

Madison, Wisconsin USA

1-608 263-5534

wiegand@cs.wisc.edu

ABSTRACT

This paper presents Semantic Web technology for the oftentimes difficult task of searching for government-produced geospatial data. We present a conceptual model that takes a task-based approach, and we formalize relationships between types of tasks, including emergencies, and types of data sources needed for those tasks. We explore the abilities and limitations of Semantic Web languages, tools, and rule engines.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Search process.
I.2.4 [Artificial Intelligence]: Knowledge Representation Formalisms and Methods - *representation languages*.

General Terms

Management, Design.

Keywords

Semantic Web, Ontology, Geospatial, Searching, Government data, Task-based.

1. INTRODUCTION

Geospatial data produced and disseminated by government agencies are important in many types of decision-making and can be vital in emergencies [2]. However, although large amounts of data are produced, potential users have a difficult time finding data sets they need. This is true despite the fact that many geospatial portals or clearinghouses are being developed by local, state, and federal government agencies. Because of the difficulties of knowing what data are available and where they might be stored, many data seekers use a general Internet search engine to search for geospatial data. Better methods are needed for effective search and dissemination of geospatial data. Furthermore, an ultimate goal would be to have a Web service automatically locate needed data sets. This possibility is becoming technically feasible using new Web technologies. The vision presented in this paper is that in an emergency, for example, a Web-based application would take in information on geographic area and type of emergency and return the types and locations of the specific geospatial data needed.

2. TASK-BASED APPROACH

2.1 Introduction

To achieve this vision, our approach formalizes the identification of tasks and their relationships to data sources. Briefly, we use OWL [1] to express ontologies that include restrictions on the types of data sources needed for each type of task. We then use a reasoner to make inferences over the knowledge base.

Our conceptual ontological model is shown in Figure 1. The main classes are task, data source, metadata, and place. Subtyping to various levels can be done for each class, and it is at the lowest/finest level that differentiations and relationships become interesting and useful. For example, a fire emergency is a type of task that may be further subdivided into particular types of fires. Tasks are related to particular data sources using OWL's restriction property. Data sources are described by metadata such as found in FGDC metadata files. Metadata contain descriptive criteria including one or more place and theme keywords and a URL to the data source. We model location as a class of places having a transitive relationship. Individuals are declared to form a test knowledge base.

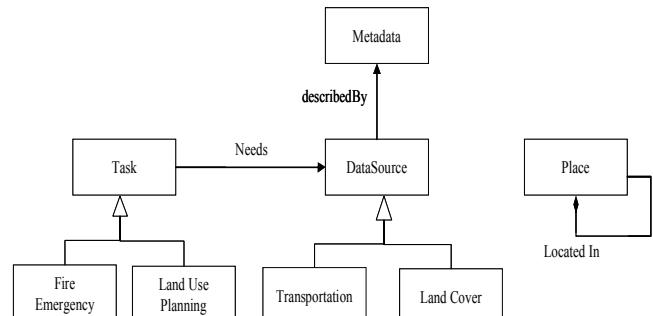


Figure 1. Conceptual model of domains

2.2 Restricting Tasks to Data Source Types

The purpose of establishing an OWL knowledge base is to allow automatic reasoning to determine data source instances needed by each type of task. We first explore conceptual design capabilities available in OWL for expressing restrictions on object properties, as was done in [4]. We then discuss the use of rule systems.

In OWL, to indicate that fire emergency tasks need certain types of data sets, we use the `someValuesFrom` property restriction in an anonymous subclass as follows:

The fire emergency class definition:

```

<owl:Class rdf:ID="Fire">
  <rdfs:subClassOf rdf:resource="#Task"/>
  <rdfs:subClassOf> {anonymous subclass}
    <owl:Restriction> {restrict to need roads, etc}
      <owl:onProperty rdf:resource="#needs"/>
      <owl:someValuesFrom rdf:resource="#Road"/>
      <owl:someValuesFrom rdf:resource="#LandCv"/>
        ...
        {other restrictions}
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>

```

We prototyped our application using the Protégé OWL ontology editor. Figure 2 shows the restriction on the *needs* object property for a fire task. The statement “ \forall needs (Road \sqcup LandCover \sqcup Hydrography)” in the middle panel indicates that all values for the needs property must be of a type listed and not of other types. The \exists statements mean that at least one child must be of that type.

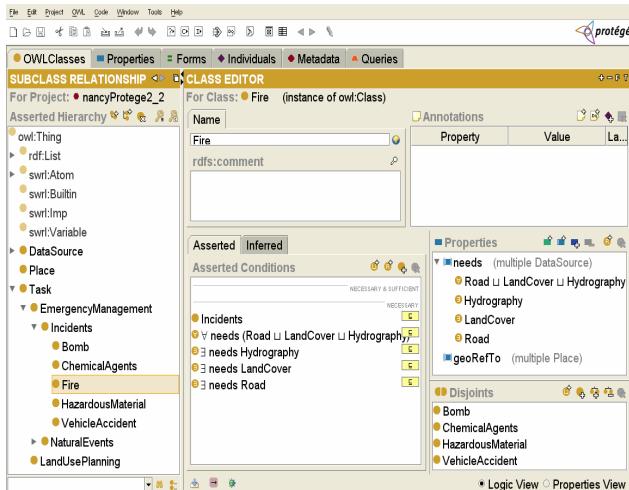


Figure 2. Screenshot of OWL ontology in Protégé.

2.3 Inferencing and Rules

We attempted to fully model the ontology such that inferencing over OWL restrictions would be sufficient to determine which geospatial data sets are needed for a task. Such an approach was used in [4] and has the benefit of maximizing the use of OWL modeling capabilities. However, we found that modeling restrictions over the numerous attributes needed for searching for geospatial data, in particular, is quite complex. Also, to search for geospatial data, comparison operators are needed to find, e.g., road data with scale equal or less than 1:24,000. As a result, we limited the modeling of restrictions in OWL to the *types* of data needed for a task. That is, similar to other researchers [e.g., 5], we needed to use rule systems to augment an ontology approach and Protégé’s limited query capabilities.

We used JessTab, a plugin to Protégé, to work with the Jess rule engine. JessTab automatically converts the OWL ontology to Jess assertions. Our prototype implementation uses fully parameterized

Jess rules that will work for any place and any type of task. Such information will be input from a Web application. Jess rules can be complex. An example is:

```

Jess> (deffrule find-data-for-task (object (:NAME ?task))(needs $?data))
  => (foreach ?element $?data (bind $?it (slot-get ?element dataKeyword)) (foreach ?element2 $?it (bind ?result (run-query* search-by-theme ?element2)) (while (?result next) (printout t "The name is " (?result getString name) " and can be found at " (?result getString url) " " crlf)))) )
TRUE

```

```

Jess> (run)

```

The name is met_CookCountyWetlands and can be found at http://www.isgs.uiuc.edu/nsdihome/browse/cook/wtldsp.e00

3. COMMENTS

Because JessTab does not convert restriction information on object properties into a Jess knowledge base, we needed to declare general instances of tasks and data sources within Protégé and then also model the restrictions using individuals. Also, because writing Jess rules is cumbersome, we are investigating SWRL, especially because SWRL has been proposed for interoperability among rule engines [3].

In summary, we used OWL, Protégé, JessTab, and Jess to create a software platform that uses ontologies, rules, and inferencing to find geospatial data. We introduced a task-based approach and formalized the relation between types of tasks and types of data sources. The full power of the approach will occur when distributed ontology tools and languages are developed. Our contribution is in exploring conceptual models and using Semantic Web tools to create a base platform from which a full expert system can be built.

4. ACKNOWLEDGMENTS

This work was partially supported by the Science and Engineering Information Integration and Informatics (SEIII) Program of NSF, Grant No. 0513605.

5. REFERENCES

- [1] Brujn, J., Polleres, A., Lara, R., and Fensel, D. OWL DL vs. OWL Flight: Conceptual modeling and reasoning for the Semantic Web. *WWW* (May 2005), Chiba, Japan.
- [2] Geospatial One-Stop, <http://www.geo-one-stop.gov>.
- [3] O’Conner, M., Knublauch, H., Tu, S., Grosof, B., Dean, M., Gross, W., and Musen, M. 2005. Supporting rule system interoperability on the Semantic Web with SWRL. *Fourth International Semantic Web Conference (ISWC2005)*, Galway, Ireland.
- [4] Stanford, How Does It Work? (OWL Example Wine Agent), <http://www.ksl.stanford.edu/projects/wine/explanation.html#ontology>.
- [5] Walker, A. Understandability and semantic interoperability of diverse rules systems, *W3C Workshop on Rule Languages for Interoperability*. (April 2005), Washington, D.C.

Design Principles for Public Safety Response Mobilization

Jane Fedorowicz

Bentley College

Departments of Accountancy
& Computer Information Systems

1.781.891.3153

jfedorowicz@bentley.edu

M. Lynne Markus

Bentley College

Department of
Management

1.781.891.2312

mlmarkus@bentley.edu

Steve Sawyer

Michael Tyworth

Pennsylvania State University
College of Information Sciences
and Technology

1.814.865.4450

sawyer@ist.psu.edu
mjt241@psu.edu

Christine B. Williams

Bentley College

Department of
International Studies

1.781.891.2655

cwilliams@bentley.edu

ABSTRACT

In this paper, we identify key characteristics of public safety response mobilization systems (PSRMS) and describe the different technical characteristics they can assume: open standards vs. commercial-off-the-shelf packages, fixed vs. mobile, security and privacy management approaches, data distribution and access control. Finally, we present testable hypotheses about which technical design features and architectural principles will produce successful PSRMS.

Categories and Subject Descriptors

[Computer applications]

General Terms

Design, Management, Standardization.

Keywords

Interorganizational systems, collaboration, public safety response, criminal justice, first responder support, design features

1. INTRODUCTION

Across the United States a new organizational form is emerging to cope with the challenges of the post 9/11 environment. This organizational form involves groups of public sector agencies and private organizations collaborating in the design and ongoing operation of complex IT-enabled infrastructures that provide ongoing support for public safety response mobilization systems (PSRMS). Three of the most extensive efforts in the US are Pennsylvania's Justice Network (JNET), the Washington D.C. area's Capital Wireless Integrated Network (CapWIN), and the San Diego, California area's Automated Regional Justice Information System (ARJIS). An example from outside of the US is the United Kingdom's Home Office and their national Police Information Technology Organization (PITO).

PSRM systems have six characteristics [cf.8; 4, 2, 10]. First,

PSRMS are *mission critical* systems. They contribute significant operational and strategic value to the organizations that use them. Second, PSRMS span organizational, geographic and political jurisdictions. Most encompass local and regional areas, while few are national at this time. Even for those with limited geographic scope, the formation and operation of these systems is complex, owing to the involvement of multiple governmental agencies (police, fire, rescue, medical services, judicial systems, human services, etc.) and private sector organizations. Third, PSRMS draw data from *multiple data repositories*. Moreover, such systems are usually designed so that host agencies maintain control of their data. Fourth, PSRMS generally support both *legacy* as well as *modern* (web-based) applications, platforms, and services. Since more than 40% of public sector workers are mobile, the systems frequently entail both *mobile* and *fixed* infrastructures and an array of devices and operating systems. Fifth, PSRMS require *sophisticated security and authentication safeguards*. Sixth, PSRMS support collaborative interaction via messaging and both *asynchronous* and *synchronous communication*.

2. INFORMATION TECHNOLOGY

The wide variety of technologies and architectures used in PSRMS demands careful attention to the ways in which their components and the combinations of components contribute to the success of PSRMS. Two dominant approaches to developing PSRMS are an open-standards approach¹ and a packaged or commercial-off-the-shelf (COTS) approach.

Table 1. Types of PSRMS and operational examples

Packaged/Vendor Design	Open Standards
COPLINK system in Fairfax County, VA; Polk County, TN	CapWIN system in DC, VA and MD metro area
Motorola Systems in Austin, TX, LA County, CA	JNET system in Commonwealth of PA
Spec. built system in Montgomery County, MD	ARJIS system in San Diego, CA area
Oracle-based system in Chicago, IL	Airwave/PITO system in United Kingdom

¹ Open standards refer to the use of publicly available technologies or their underlying policies/protocols/standards, including such things as XML, browser standards (http, html, etc.), and published APIs.

PSRMS also can be characterized by their other technical characteristics: fixed vs. mobile, approach to security and privacy management, data distribution and access control. Each of these technical characteristics involves tradeoffs that must be considered in the context of PSRM design. A careful study of what makes PSRMS successful must draw on the extensive literatures on technologies and system engineering and must pay careful attention to technical characteristics (and support issues) and factors contributing to success of these interorganizational systems [3, 5, 6, 9].

3. TESTABLE RESEARCH QUESTIONS

The literature on interorganizational information systems suggests that some technical design features and architectural principles will produce more successful PSRMS than others. For example,

1. PSRMS are more likely to be successful when based on open standards than on vendor-specific standards.

Each approach has strengths and weaknesses [8]. Open standards can reduce upfront investment costs, which are important in a public sector environment, because they enable agencies to make use of existing, standards-based technology. The costs for systems integration and ongoing support can be greater, however, because these solutions may require more technical expertise at participating organizations. On the other hand, although proprietary solutions may involve higher conversion costs, they may exhibit lower "out of the box" integration costs and may provide better third-party support.

2. PSRMS are more likely to be successful if they are integrated with participants' existing back-end systems.

Lack of integration between interorganizational systems like PSRMS and the participating organizations' back-end systems has been shown to be a major source of problems including lack of adoption of the systems; low levels of use or ineffective uses; errors, delays, and missing data; and lack of benefits [1]. Although integration can be costly, most analysts identify it as a critical success factor for interorganizational systems.

3. PSRMS are more likely to be successful if they rely on federated data sources rather than on centralized data sources.

When PSRMS rely on centralized data repositories, these repositories do not necessarily replace local databases. Local participants may choose not to supply data to a central repository, fearing loss of control and the extra costs of maintaining a parallel system. In addition, if local participants do not implement two-way integration with the centralized data resource, the problems described earlier can result.

4. PSRMS are more likely to be successful if there are commercial, off-the-shelf solutions for specific needs.

PSRMS must experience a high degree of acceptance across all participating constituencies in order to achieve success. For some small or technologically unsophisticated participants, COTS systems may be seen as easier and cheaper to acquire. In other cases, COTS components are preferred when including commodity-like functionality, such as VoIP or Instant Messaging for ad hoc team communications.

5. PSRMS are more likely to be successful if there is technical support for user authentication and security protection.

A common problem for first responder organizations is limited IT support and diverse IT infrastructures [7]. Without technical support, officers rely on themselves and each other to troubleshoot problems. Gil-Garcia [2] found that third-party technical assistance helped ensure success for justice initiatives. Third-party support is particularly needed in the areas of user authentication and security protection. However, the technical knowledge and skills required to ensure information security are not readily available in small and public sector organizations.

Public sector and PSRM IT-related initiatives in particular, may encompass different definitions of or criteria for what constitutes success. Testing of the hypotheses presented in this paper will require measures of success to be defined and analyzed within the context of the PSRMS under study.

4. ACKNOWLEDGMENTS

Support for this work provided by National Science Foundation grants IIS-0534877, IIS-0534889, IIS-051248, and the Boston University Institute for Leadership in a Digital Economy.

5. REFERENCES

- [1] Barua, A., Konana, P., and Whinston, A.B. "An Empirical Investigation of Net-Enabled Business Value," *MIS Quarterly* 28, 3 (2004) 558-620.
- [2] Gil-Garcia, J. R., Schneider, C. A., and Pardo, T. A. Effective Strategies in Justice Information Integration. Center for Technology in Government, State University of New York, Albany, NY, 2004.
- [3] Lin, C., Hu, P. J.-H., and Chen, H. Technology Implementation Management in Law Enforcement: COPLINK System Usability and User Acceptance Evaluations. *Social Science Computer Review* 22, 1 (2004), 24-36.
- [4] National Association of State Chief Information Officers (NASCIO). Concept for Operations for Integrated Justice Information Sharing Version 1.0. Retrieved July, 2004 from <http://www.nascio.org/publications/index.cfm>.
- [5] Northrop, A., Kraemer, K. L., and King. Police use of computers. *Journal of Criminal Justice*, 23, 3 (1995), 259.
- [6] Nunn, S. Police information technology: Assessing the effects of computerization on urban police functions. *Public Administration Review*, 61, 2 (2001), 221.
- [7] Sawyer, S., Tapia, A., Pesheck, L., and Davenport, J. Mobility and the First Responder. *Communications of the ACM*, 47, 3 (2004), 62-65.
- [8] Sawyer, S., Allen, J., and Lee, H. Broadband and Mobile Opportunities: A Sociotechnical Perspective. *Journal of Information Technology*, 18, 1 (2003), 121-136.
- [9] Turoff, M., Chumer, M., et al. Assuring Homeland Security: Continuous Monitoring, Control and Assurance of Emergency Preparedness. *JITTA* 6, 3 (2004), 1-24.
- [10] Williams, C. B. and Fedorowicz, J. A Framework for Analyzing Cross Boundary e-Government Projects: The CapWIN Example. *Proceedings of the National Conference on Digital Government Research* (Atlanta, May, 2005).

University Information System RUSSIA: data, knowledge products and services for social research

Tatyana Yudina,

Leading researcher, Ph.D.

Moscow State University Research Computing Center,
Vorobiovy Gory, Moscow, Russia, 119992

7 495 939 3015

yudina@mail.cir.ru

Categories and Subject Descriptors

H.2.4 ORACLE, H.2.4 SQL Server , D.3.2 JavaScript

General Terms

Management, Design, Economics.

Keywords

Information system, content-based search engine, informers, value-added services, relational database, system analysis, social research, knowledge products, bilingual search.

1. MISSION AND CONTENTS

Moscow State University - based University Information System RUSSIA (UIS RUSSIA, www.cir.ru) has been designed and is maintained as a digital library for research and education in economics and social sciences. It has been in operation since 2000.

The system maintains collections of social domain data and documents. Currently two million documents from 60+ collections are integrated. Contents include official data and documents (laws, presidential decrees and directives, governmental enactments, acts and regulations); international agreements signed by the Russian Federation (RF); stenograms (daily records) of State Duma of the Federal Assembly of RF; state statistics; analytical reports of government agencies; analytical journals; reports and databases maintained by "think tanks"; academic publications – Moscow State University Bulletin (social sciences series), "Sociological Journal", "Forecast", DemoscopeWeekly, etc.; public opinion polls data; mass media. Holdings in English are also maintained - archive of academic publications in economics and social sciences (full text documents available in RePEc database); international organizations documents and databases; foreign universities collections.

2. NLP TECHNOLOGY. SEARCH ENGINE

In order to maintain the UIS RUSSIA as an integrated resource with content-based searching across collections technology for automatic linguistic text processing (ALTP, special software-lingware-knowledgeware complex) has been designed, developed and implemented within the framework of the project. The NLP technology is adjusted to process all main types of business prose text corpora.

The procedures include processing of electronic text in several main formats (ASCII, HTML, MS Word) in Windows and operating as DLL; morphological analysis of Russian/English texts; terms' recognition/disambiguation; Thesaurus-based thematic analysis—event categorization, indexing, annotation/summarization; download of results to an Oracle database server. The main ALPT instrument is a Socio-Political Thesaurus (Thesaurus). Its current version incorporates 70,000 concepts/descriptors with synonyms, including 6,500 geographic names. Thesaurus is bilingual. The tool assists in identifying main and subordinate topics in a document. The NLP technology and bilingual Thesaurus provide for processing and annotation in Russian of documents in English. The technology provides for up to 500 Mb of electronic texts to be processed and integrated into the University Information System RUSSIA daily. ALPT results are utilized to ensure advanced search engine - on top of traditional tools - content-based search and query refinement is available, exploiting Thesaurus with thesaurus hierarchy-based query refinement; several systems of subject headings, including that of the UIS RUSSIA and Congressional Research Service, Library of Congress, USA, Legislative Indexing Vocabulary Top Terms; JEL (Journal of Economic Literature)-based classification system.

3. VALUE-ADDED SERVICES FOR ASSISTING RESEARCH

The UIS RUSSIA provides for value-added services to assist research. The most developed are the services for state statistics-based investigations. With the Russian state statistics agency publishing data in aggregated tables and mostly in .doc format the following UIS RUSSIA services are most important :

- state statistics converted into relational data base format (power tables) available at federal, regional and local levels;
- MS Excel 97 format available for all statistical tables, including the tables presented in analytical reports and scientific journals;
- Links to the Methodological Notes and Glossary for statistics;
- System of Subject Headings to integrate data from different publications;

- Content-based search exploiting Thesaurus and subject domain ontologies.

4. SUBJECT-ORIENTED MODULES

Subject-oriented modules are implemented to cover the most demanded domains of social investigations and university education courses. The modules integrate data and knowledge products and provide for monitoring and analysis of economic, fiscal, demographic, agrarian tendencies and comparison at regional and local levels. UIS RUSSIA Thesaurus assists in cross-search and navigation. To refine search technique and knowledge management in public finances domain a subject-oriented ontology is under construction. Ontology is bilingual to provide for also documents in English retrieval.

In 2005 relational database is implemented to integrate economic, social and budget data and provide for interdisciplinary and system investigations. GIS elements are available to provide for map-based data presentation and analysis. The table below presents the data on budget funding for theaters and other performing arts organizations in 2002 (published by Ministry of Finance) and number of persons attended the performances in 4 regions of RF (data is published by Russian State Statistics Agency). Data from other state agencies will be added in 2006. Contents, technology and value-added user services make the UIS RUSSIA a valuable resource for full-scale interdisciplinary and socially relevant investigations.

The system is free for researchers and educators, registration is needed. 400+ universities, higher education institutions, colleges, academic institutes, think tanks and 4000+ individuals are subscribed and work with the system. The system is also accessible via public libraries and assists in citizens education. RF state agencies of federal, regional and local levels are becoming active in exploiting the UIS RUSSIA services. Workshops and training are regularly arranged to assist in government agencies staff education in data analysis in order to stimulate e-government technologies and principles in Russia.

5. ACKNOWLEDGMENTS

Since 1993 the project has been supported by grants from the Russian Fund for Basic Research (№ 0407 90279B); the Russian Fund for Humanities (№ 0402-12004/B); the Ministry of Science and Technologies of RF's "Informatization of Russia" program; the MacArthur Foundation, USA; the Ford Foundation, USA; and the Eurasia Foundation, USA.

Шаг 1. Выбор отчета	Шаг 2. Выбор территории	Шаг 3. Выбор года	Шаг 4. Выбор план/исполнение	Шаг 5. Выбор показателя	Шаг 6. Конструктор	Шаг 7. Таблица
Сводная таблица						
		2002				
		Исполнено по бюджету субъекта Российской Федерации.				
		Государственная поддержка театров, концертных организаций и других организаций исполнительских искусств.		Численность зрителей театров на 1000 человек населения, человек		
Ростовская область	65141			118		
Республика Башкортостан	217189			199		
Челябинская область	77198			163		
Иркутская область	108896			157		
© Регионы России. Основные характеристики субъектов Российской Федерации. 2004: Стат.сб./Госкомстат России. - М., 2004.						
© Отчеты об исполнении консолидированных бюджетов субъектов РФ и муниципальных образований за 1998 - 2004 гг. /Министерство финансов РФ						

Figure 1. Database on Social and Budget Statistics of Russia

Connected Kids: Designing a Youth-Services Information System for Local Government

James P. Zappen
Rensselaer Polytechnic Institute
Department of Language, Literature,
and Communication
Troy, New York
518-276-8117
zappenj@rpi.edu

Sibel Adali
Rensselaer Polytechnic Institute
Department of Computer Science
Troy, New York
518-276-8047
sibel@cs.rpi.edu

Teresa M. Harrison
University at Albany, SUNY
Department of Communication
Albany, New York
518-442-4883
harrison@albany.edu

ABSTRACT

The Connected Kids research team has launched a working prototype of its youth-services information system and has further developed the prototype, adding customizable web pages, a document-entry function, and a new map interface and initiating development of additional search capabilities. To promote use of the system, we have provided support for registration and data-entry activities, launched a publicity campaign, and conducted system evaluations. To ensure long-term sustainability of the system, we have extended our registration and data-entry activities beyond our local community, continued our negotiations with our government partner, and initiated discussions with other government units and organizations.

Categories and Subject Descriptors

J.4. [Computer Applications]: Social and Behavioral Sciences –
Communication

General Terms

Design, Theory, Experimentation

Keywords

Digital Government, City Government, County Government,
Information System, Youth Services

1. PROJECT BACKGROUND

Connected Kids is a youth-services information system for Troy and Rensselaer County, New York, currently accessible as a working prototype (<http://www.connectedkids.info/>, January 12, 2006). The information system is a collaborative venture between Rensselaer Polytechnic Institute and the City of Troy, in cooperation with Rensselaer County, the Troy and Lansingburgh public schools, and more than twenty youth-services organizations.

The information system includes a database of information about programs and activities for families and children, with sophisticated search capabilities, a distributed data-input function, and separate interfaces for parents, teens, and children; galleries of children's artwork and photos; and a distribution system that extends the services of the system to low-income families at the Troy Housing Authority. Such a system requires sophisticated

search technology and interface design and a social network to support registration and data entry, publicity, and long-term maintenance.

Within the past year, Connected Kids has extended its collaborations to include an active Advisory Board and additional organizational participants; launched a working prototype and added new features; and initiated a successful publicity campaign and system evaluation. Our challenges include more consistent data entry by organizational participants and long-term sustainability of the system. Our most promising opportunities lie, we believe, in enhanced search functionalities that have the potential to substantially increase the data accessible via the system and in continued outreach to additional government and organizational partners within the immediate area and the larger Capital Region.

2. GOALS AND ACCOMPLISHMENTS

Collaborations with government and organizational partners support the continued development of the system prototype and promotional activities such as data entry, publicity, and system evaluations.

2.1 Collaborations with Government and Organizational Partners

Collaborators include the City of Troy, an active and supportive Connected Kids Advisory Board, and a growing list of organizational participants. In the past, the City has provided consultation on system design, direct support, and cost sharing. Currently, the City provides a representative to serve on the Advisory Board and supports ongoing publicity efforts. The Advisory Board provides regular consultations on system registration and data entry, publicity, and long-term maintenance issues. Organizational participants include twelve members of the Advisory Board and approximately fifty current system registrants. Several members of the Advisory Board experiment regularly with the system, produce novel features that can be incorporated into the system, and identify new needs and opportunities for additional system enhancements.

2.2 Development of the System Prototype

Recent developments of the system prototype include customizable web pages with a document-entry function, a new map interface, initial development of enhanced search

capabilities, and enhancements to the dissemination system and the art and photo galleries. The customizable web pages permit organizations to create their own World Wide Web presence and to post .doc or .txt documents, which are automatically converted to .pdf files and linked to these pages.

Other developments utilize technologies such as the Google API to enhance system functions (<http://www.google.com/apis/>, January 12, 2006). The new map interface, now in beta, permits users of the system to select an organization name or a point on a map and thereby access information by pointing and clicking rather than conventional searching or browsing. We are in the process of extending this interface with a more extensive search capability. We are also investigating the possibility of extending the search capability of the system to the webs of all registered organizations, including large webs, such as government or school webs, thereby greatly enhancing the quantity and quality of information accessible via the system, without additional data entry by organizational participants.

The dissemination system permits the Troy Housing Authority to distribute application forms, newsletters, and other information via its customizable web pages at four sites that currently have Internet connections, with more forthcoming as the connections become available (<http://www.connectedkids.info/tha/>, January 12, 2006). The galleries display slideshows of children's artwork and photos and now permit organizational users to create their own background images, which highlight their logos, color schemes, and other aspects of their organizational identities.

2.3 Promotional Activities

Promotional activities include support for registration and data entry, a publicity campaign, and ongoing system evaluations.

Support for the registration and data-entry activities includes regular email solicitations, data-entry sessions at one of our on-campus research labs, and on-site visits to organizational participants by members of the Connected Kids research team.

Publicity efforts include mass mailings of Connected Kids postcards, poster displays, and distribution of bookmarks to local school children (Figure 1. Connected Kids Bookmark).



Figure 1. Connected Kids Bookmark

These efforts also include displays and distribution of publicity materials at street fairs and block parties, community walks/runs and health fairs, the local Farmer's Market, and the most heralded of Troy events, the Victorian Stroll. Results of these efforts show heavy increases in traffic to the Connected Kids system, with spikes exceeding 2,500 hits per day in May, August, and November, 2005.

Ongoing system evaluations include sessions with both children and parents at the Troy Boys & Girls Club. Preliminary

assessments of the results show positive responses to the system and especially the galleries but underscore the need for regular and systematic data entry to keep the system live and active with information about current activities and programs for families and children.

3. CHALLENGES AND OPPORTUNITIES

Challenges include both increased registration and data entry and long-term sustainability of the system. These challenges might be addressed by continued support for the registration and data-entry activities and enhanced search capabilities and by the enlistment of additional government and organizational partners.

3.1 Registration and Data Entry

The data-entry procedure, via a simple, easy-to-use, copy-and-paste system function, does not appear to be prohibitively difficult in principle, but in practice it seems to require constant monitoring and attention by both the Connected Kids research team and the organizational participants. An enhanced search capability, utilizing a technology such as the Google API referenced above, might help to alleviate this difficulty to some degree by permitting users to search any and all webs linked to the Connected Kids system, whether or not the organizational participants have entered the contents of their webs as individual activities or programs. At the same time, the use of this technology would not preclude the entry of these individual items of information.

3.2 Long-Term Sustainability

The long-term sustainability issue is currently our most pressing issue since it is dependent upon a commitment from a government partner to maintain the information system on its own server and network connections for the indefinite future. As a response to this issue, we are maintaining our longstanding partnership with the City of Troy, and we are also reaching out to other government units and organizations within the three-county Capital Region, both as potential system registrants and potential partners. This outreach, of course, addresses both the data-entry and the sustainability issues since it increases the amount of data potentially accessible in the system and the number of potential government and organizational partners. In addition, this outreach also extends the potential use of the system to the entire region, not just Troy and Rensselaer County.

We remain optimistic, therefore, that new developments in search technology and the institutional resources available to us within the larger Capital Region will provide opportunities to address the challenges of operating and sustaining the system as a valuable resource for families and children within both our local community and our larger three-county region.

4. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 0091505. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Accuracy Improvement of Urban True Orthoimage Generation Using 3D R-tree-based Urban Model

Guoqing Zhou and Wenhan Xie

Laboratory for Earth Observing and Informatics

Old Dominion University, Norfolk VA 23508

Email: gzhou@odu.edu, wxie@odu.edu

ABSTRACT

When applying the existing TRUE orthorectification methods to orthorectify urban high buildings, there exist many problems such as incomplete orthorectification of boundaries of artificial buildings, edge blurring and distortion; incomplete orthorectification to small buildings in building roofs, and/or balcony due to the lack of proper 3D model representation. This paper presents a novel orthorectification method that is based on 3D R-tree CSG urban building model. This method can exactly solve the problem mentioned above and improve the accuracy of true orthoimage. A test field located in downtown of Denver, Colorado has been used to test our methods. The experimental results demonstrated that the proposed method in this paper can effectively improve the accuracy of urban buildings with 3-5 pixels, especially for the boundaries of building, and some small buildings can completely orthorectified.

Keywords

True orthoimage, DBM, orthorectification, 3D R-tree, CSG

1. INTRODUCTION

Digital orthoimages are a critical component of the National Spatial Data Infrastructure (NSDI) [15]. In recent years, with the increasing importance of GIS and aerial photogrammetry, the task of the urban planning and developing gradually rely upon the true orthophotos. Many efforts about true orthophoto generation have been made, e.g.,[1][2][12][15]. Theoretically, the digital orthoimages should be a spatially accurate image with ground features represented in their true planimetric positions [15]. However, the algorithms and procedures in traditional digital orthoimage generation did not consider the spatial objects, such as buildings, resulting in the spatial objects of orthoimage in urban areas are distorted from their true positions. This distortion shows that buildings lean over a street, thereby, occlude the street and other artificial buildings.

Due to the problems mentioned above, this paper mainly studies the accuracy improvement of the true orthophoto generation. In order to orthorectify a building to its correct, upright position, an exact digital building model (DBM), which describes the building structure, three-dimensional coordinates, topologic relationship, etc., is required [15]. Aiming at this purpose, the research about automatic or semiautomatic building extraction [9][14][6] become a key problem. Many different approaches and algorithms were addressed with different source image type, terrain complexity and field of application etc. In recent decade

years, CSG model has commonly used for building extraction in the field of computer vision and photogrammetry [3][10][5][13], because of its flexibility for the representation of buildings and its diversity for containing object constraints and classification.

2. CSG-BASED BUILDING MODEL EXTRACTION

As described above, DBM-based orthoimage generation requires an effective representation of urban buildings. The generation of a 3D urban building model is a rather challenging task because different applications require different data types and manipulation functions. Many models have been proposed, e.g. [4][7][8]. For the purpose of true orthoimage generation, the data structure to be developed in this paper requires not only the fitness for generating the DBM-based high-quality orthoimage, but also easily creating, storing, designing, analyzing and querying city objects for orthoimage-based urban applications.

This paper constructs the urban building model using CSG model organized by 3D R-tree. The main characteristics of the classical R-tree is summarized as a collection of tuples, each tuple has unique identifier, leaves contain a tuple of the minimum-bounding rectangle (MBR) of an object and non-leaves contain another tuple of the MBR of its lower entries[16]. In this paper, each CSG model is composed of a combination of volumetric primitives with 3D minimum bounding box. It is possible to construct a complex model with a small set of primitives, depending on the level of detail required. A primitive is a predefined simple solid model to determine the intrinsic geometric properties of a building, and is associated with some transformation parameters to perform scaling, rotation, and translation. The combination of primitives can finish via Boolean set operations, such as union, intersection, and difference. With 3D R-tree organization, we can produce the dynamic 3D digital building model (DBM) with varying levels of detail, which is range from tiny parts of one building to the whole block of street.

The process of extracting building knowledge is a bottom-up procedure from low-level data to high-level information. The workflow is illustrated by Fig. 1(a). First, we can obtain the vector edge data of buildings from aerial images or LIDAR data by using a series of image processing such as edge detecting, tracking and segmenting etc. after feature grouping, the upper-level building outline is generated. As we know, most of architectures are with regular shapes, where the lines and planes on them have the properties of parallel, vertical, orthogonality etc. With these geometric constraints these outlines can be matched with corresponding CSG primitives in the database of primitives. These CSG primitives are parameterized by the

building heights from digital surface model. For the complex buildings, they are decomposed into several parts. We can combine a complete building model via Boolean set operations, which include union (\cup), intersection (\cap), difference ($-$), etc. Boolean operations make the CSG model more flexible. For example, the redundancy in space may be occurred for complex building where some parts cover the others in the ground map. In this case, the

Boolean difference operated can reduce the data redundancy. With the connection of Boolean operations, the whole composition process of a complex building model likes a CSG tree (Fig. 1(b)). The leaf nodes are those CSG primitives; the middle node links two branches of combined parts and the root is the complete building model.

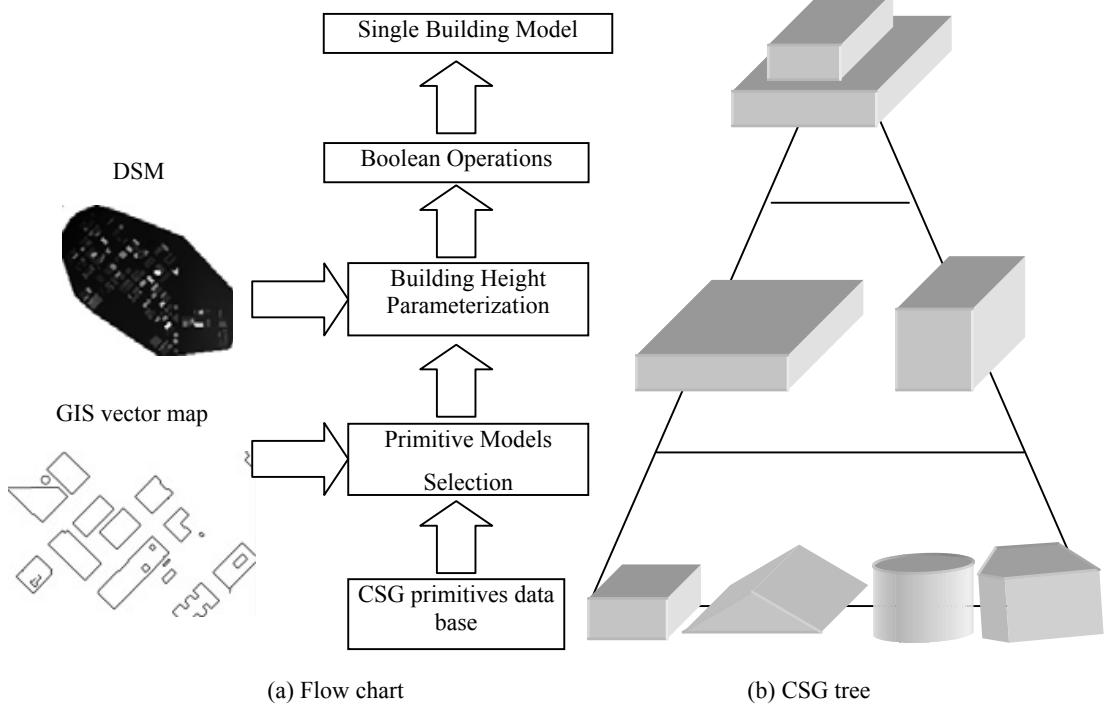


Figure 1: Procedure for extracting CSG building model

3. ACCURACY IMPROVEMENT

In conventional orthorectification, exterior orientation parameter is derived from space resection with control points. However, in urban area, the phenomenon of occlusion among the buildings or themselves often happens. Because of the occlusion of corresponding control points, the control points are distributed unevenly, thus it results that the accuracy of algorithm is not stable enough. In order to improve the accuracy of orthorectification, the relative and absolute control information appeared in the reference object should be considered. These control information includes the distance (between the central point of camera and 3D space point); the point located in a plane; the point located in a straight line; the known angle and other known graphs.

In building models, there exists plenty of geometric information in the façade of building, such as perpendicularity, parallelism among straight lines and planes. All of these relationships can be regarded as the constraint conditions of unknown parameters (i.e. relative control conditions).

As we known, the collinear equation of photogrammetry is

$$x - x_0 = -f \frac{a_1(X - X_s) + b_1(Y - Y_s) + c_1(Z - Z_s)}{a_3(X - X_s) + b_3(Y - Y_s) + c_3(Z - Z_s)}$$

$$y - y_0 = -f \frac{a_2(X - X_s) + b_2(Y - Y_s) + c_2(Z - Z_s)}{a_3(X - X_s) + b_3(Y - Y_s) + c_3(Z - Z_s)}$$

The general matrix form of its error equations is

$$V = AT + BX - L$$

where, T is the matrix of orientation parameters; X is the matrix of the coordinates of object points.

$$T = (\Delta f \Delta x_0 \Delta y_0 \Delta \varphi \Delta \omega \Delta \kappa)^T$$

$$X = (\Delta X_1 \Delta Y_1 \Delta Z_1 \dots \Delta X_i \Delta Y_i \Delta Z_i \dots)^T$$

The geometric constraints can be found in the information of building model:

$$C\hat{X} + W_x = 0$$

Combining two formulas, the adjustment model is:

$$\begin{cases} V = AT + BX - L \\ C\hat{X} + W_x = 0 \end{cases}$$

The equation below is for computing the optimum parameter value:

$$\Phi = V^T PV + 2K_s^T(C\hat{X} + W_x)$$

The normal equation is:

$$\begin{cases} B^T PB \delta X + C^T K_s + B^T PL = 0 \\ C \delta X + 0K_s + W_x = 0 \end{cases}$$

Let $N_{bb} = B^T PB$, the expression above can be rewritten:

$$\begin{pmatrix} N_{bb} & C^T \\ C & 0 \end{pmatrix} \begin{pmatrix} \delta X \\ K_s \end{pmatrix} + \begin{pmatrix} B^T PL \\ W_x \end{pmatrix} = 0$$

Suppose that

$$\begin{pmatrix} N_{bb} C^T \\ C \ 0 \end{pmatrix}^{-1} = \begin{pmatrix} Q_{11} Q_{12} \\ Q_{21} Q_{22} \end{pmatrix},$$

The solution of parameters can be finally calculated:

$$\begin{cases} \delta X = -(Q_{11} B^T P L + Q_{12} W_X) \\ K_S = -(Q_{21} B^T P L + Q_{22} W_X) \end{cases}$$

According to these geometric constraints, this paper addresses CSG model feature-based orthorectification to improve the accuracy of true orthoimage. This method fully utilizes the geometric relationships of 3D model vector feature existing the buildings, thus the control information is not only ground control point, but also relative control information. In this paper, two different types of relative control information are considered:

- 1) Object-based control information; there are inherent geometric characters in the façade of building, for example, the vertical line of building is perpendicular to the edges of roof, and to those rectangular roofs, the edges of roof are with the same height and are orthogonal each other. According to the special data structure mentioned in Section 2.1, each CSG model is with topological relationship, therefore, object-based relative control condition can be inherited automatically from the topology of CSG model.
- 2) Image-based control information. Theoretically when CSG models are back projected into the original aerial image, the projected model should be corresponded with the outline of building in aerial image. However, due to various systematic and occasional errors, they can't be exactly matched. Therefore, we can deductive the geometric constraints between the features projected from CSG building models and the edges of buildings extracted from aerial image.

Both of these control information, which are based on feature, needn't any control point information.

For object-based geometric constraint, this paper only takes the most common geometric relationship for example: perpendicularity and collinearity between two segments.

Suppose that segments ab and bc are two edges of one building roof in the original aerial image, the arbitrary image point a, b and c are as observation values. Both of two segments correspond with the space segments AB and BC respectively. Segments AB and BC are with the same height (corresponding with Z-axis in space coordinate system), which is derived from CSG building model.

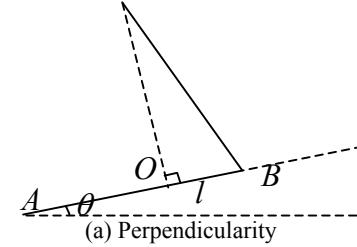
Perpendicular constraint: Shown as Fig. 2(a), suppose that segments AB and BC are perpendicular each other at joint point B. Due to the same height in one roof, we can calculate the coordinate of point A, B and C in plane X-Y, which are expresses as (X_A, Y_A) , (X_B, Y_B) and (X_C, Y_C) . The angle θ is the angle between AB and X-axis. Point O is the orthocenter from C to segment AB. L is the distance of B and O, it can be expressed as the formula as below:

$$l = (X_C - X_B) \cos \theta + (Y_C - Y_B) \sin \theta = (X_C - X_B) \frac{(X_B - X_A)}{S_{AB}} + (Y_C - Y_B) \frac{(Y_B - Y_A)}{S_{AB}}$$

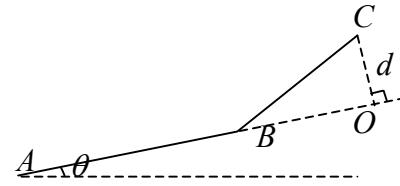
Where S_{AB} is the distance of segment AB.

Theoretically the distance l should be zero. There is the equation with the differential form:

$$\begin{aligned} l_0 + \Delta l &= [(X_C - X_B) \frac{(X_B - X_A)}{S_{AB}} + \\ &(Y_C - Y_B) \frac{(Y_B - Y_A)}{S_{AB}}] + \frac{1}{S_{AB}} [(X_B - X_A) \Delta X_C + \\ &(Y_B - Y_A) \Delta Y_C + (X_B - X_C) \Delta X_A + \\ &(Y_B - Y_C) \Delta Y_A + (X_C - 2X_B + X_A) \Delta X_B + \\ &(Y_C - 2Y_B + Y_A) \Delta Y_B] = 0 \end{aligned}$$



(a) Perpendicularity



(b) Collinearity

Figure 2 two types of object-based geometric constraints

Collinear constraint: shown as Fig. 2(b), suppose that segments AB and BC are collinear. The definition of θ and O are the same as above. Distance d is the distance of C and O, and it can be expressed as the formula as below:

$$d = (Y_C - Y_B) \cos \theta - (X_C - X_B) \sin \theta = (Y_C - Y_B) \frac{(X_B - X_A)}{S_{AB}} - (X_C - X_B) \frac{(Y_B - Y_A)}{S_{AB}}$$

Where S_{AB} is the distance of segment AB.

Theoretically the distance d should be zero. Therefore, the equation with the differential form is:

$$\begin{aligned} d_0 + \Delta d &= [(Y_C - Y_B) \frac{(X_B - X_A)}{S_{AB}} - \\ &(X_C - X_B) \frac{(Y_B - Y_A)}{S_{AB}}] + \frac{1}{S_{AB}} [(Y_A - Y_B) \Delta X_C + \\ &(X_B - X_A) \Delta Y_C + (Y_B - Y_C) \Delta X_A + (X_C - X_B) \Delta Y_A + \\ &(Y_C - Y_A) \Delta X_B + (X_A - X_C) \Delta Y_B] = 0 \end{aligned}$$

There are the similar differential equations as above in the case that all of these segments locate on plane X-Z or plane Y-Z. If point a, b and c are corresponding with ground control points (GCP), we can get three observation equations and one condition equation. For perpendicular and collinear constraints, each condition equation will introduce six unknown parameters to be solved.

For image-based geometric constraint, suppose that the feature in aerial image can be depicted as the quadratic equation $\mathbf{x}^T \mathbf{Q} \mathbf{x} = 0$, where vector $\mathbf{x} = (x, y, 1)$ and vector \mathbf{Q} is constant matrix. The projection of the corresponding portion of CSG model will match with this feature. We take the some projection points \mathbf{x}_i as virtual observation values. Its ideal value $\hat{\mathbf{x}}_i$ fit the equation

$\hat{\mathbf{x}}_i^T \mathbf{Q} \hat{\mathbf{x}}_i = \mathbf{0}$. After linearization, we can get the condition equation. To simply the complexity of algorithm, according to specificity of urban area, this paper only considers the linear equation.

All of these absolute and relative control conditions can combine together into the geometric feature constraint based bundle adjustment. Usually the corners of the buildings are selected as control points, and the features, which are composed of these control points, also can be used as constraint conditions.

The workflow of accuracy improvement of urban true orthoimage generation using 3D urban model is illustrated in Fig. 3. First, the initial value of orientation parameters is calculated by space resection from digital building model and aerial image. According to these orientation parameters, we project the CSG building models back into original aerial image. Thus preliminary outlines of buildings with vector feature are respectively framed to the corresponding 2D building image. For each building in original aerial image, vector features can be extracted automatically, these features are matched with projected building model frame. For architectures, most of building roofs are rectangular or regular shapes, and vertical with upright wall outlines. These can be taken into account as geometric constraint and added to control point-based bundle adjustment algorithm. This algorithm can iteratively adjust orientation parameters, consequently the true orthoimage can be generated more accurately by using CSG model feature based orthorectification.

4. EXPERIMENTS AND ANALYSES

The experimental field is located in downtown Denver, Colorado, where the highest building is 125m, and many others are around 100m. The six original aerial images from two flight strips were acquired using an RC30 aerial camera at a focal length of 153.022 mm on April 1, 2000. The flying height was 1650m above the mean ground elevation of the imaged area. The aerial photos were originally recorded on film and later scanned into digital format at a pixel resolution of 25 μm . One part of scanned aerial images is shown as Fig. 4. Fig. 5 is the 2D representation of the digital surface model of the same area.



Figure 4: Part of original aerial images

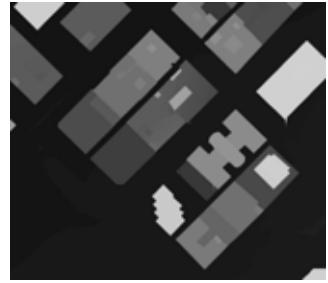


Figure 5: digital surface model in downtown Denver used for the experiment. The brightness in the DSM represents surface height

From the digital surface model, we extract 1386 edge objects including point, line and arc objects. After feature grouping, 249 building parts is used for 2D GIS vector map where every object has the property of height. According to 2D geometric information and 3D height information, all the GIS vector objects can be matched with parameterized CSG primitives. In this experiment, by combining CSG primitives, 106 buildings are finally extracted. Fig. 6 shows the partial extracting result compared with corresponding aerial image.

Aerial image	Corresponding building model

Figure 6: The extracting result compared with corresponding aerial image

The interior orientation parameter is known: the focal length is 153.022mm; the principal point is (0.002mm, -0.004mm). Firstly, we select eight random control points to calculate the initial value of the exterior orientation parameter, which is listed in 1st row of Table 1. The result of the exterior orientation parameter derived from traditional orthorectification method based on 331 control points is listed in 2nd row of Table 1. These control points are commonly selected at the corners of building bottoms and roofs.

With these initial values, 106 CSG building models (model frame is shown as Fig. 7) is back projected into original aerial image. One of the projected model frames (called building 1 in this paper) is shown as Fig. 8. Note that the projected model frames are rendered with transparent mode, so the occluded parts of frames can be seen. From Figure 8, CSG building frame is not exactly located the appropriate

position. There are dozens of pixel's difference during orthorectification. Fig. 9 is the result of line extraction of this building. All of these extracted lines are considered candidate lines to be matched with projected building model. Fig. 10 shows the matched lines marked with red color. According to the improvement algorithm based on

bundle adjustment of geometric constraints, improved orientation parameter is obtained. The result of orientation parameter improvement is listed in the last row of Table 1. Fig. 11 is the result of projected CSG frame, whose parameters are bundle adjustment mode constrained by building features.

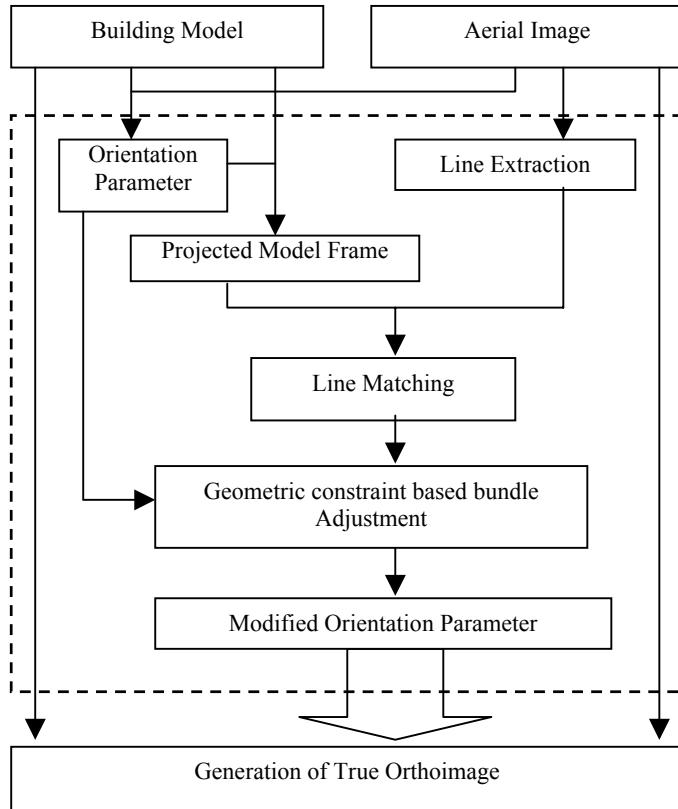


Figure 3: Flowchart about accuracy improvement of true orthoimage



Figure 7: 3D building frame

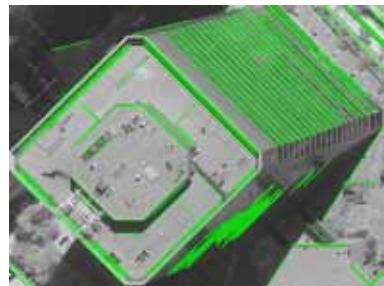


Figure 9: Line extraction



Figure 8: Projected CSG model frame of building one

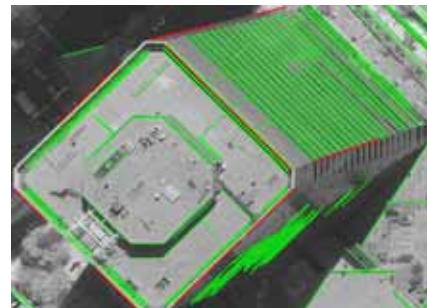


Figure 10: Line matching with model frame

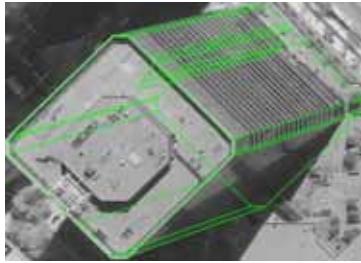
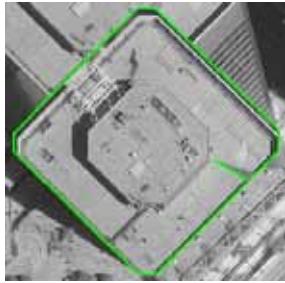


Figure 11: The result of projected CSG frame using the improved orientation parameter obtained from bundle adjustment model, which building features are used as constrain.

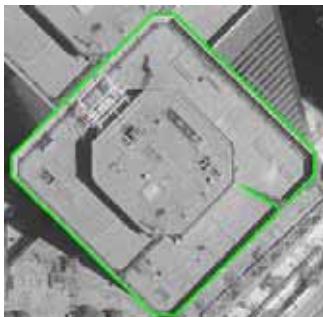
From the results above, we can see the roof of the projected building model frame is exactly settled with images.

In this test area, there are plenty of line features, after line extraction and matching, 76 object-based and image-based line constraints together with 331 control points are added to the bundle adjustment system. After iterative calculation, the RMS of system $\sigma_0 = 0.0346$.

The final result comparison of building one's orthorectification before and after accuracy improvement is shown as Fig. 12. Fig. 12(a) is the result of before accuracy improvement. Fig. 12(b) is the result of the method presented by this paper. The green-color outline is the right position of the building roof in the orthophoto. We can see that there's obvious offset in both x and y direction away from its right position in 12(a), while with CSG model constraint, the orthophoto of building one almost exactly locates its right position.



(a) The orthorectification result before accuracy improvement



(b) The result after accuracy improvement

Figure 12: the comparison of building one's orthorectification result before and after accuracy improvement

5. CONCLUSION

In this paper, we developed the accuracy improvement of true orthophoto. With 3D R-tree organized CSG building model, we presented a novel accurate improvement method of exterior orientation elements based on the building's features, which are used for constraints of bundle adjustment. These exterior orientation elements are used for generation of true orthoimage, thus, it is expected to improve the accuracy of orthoimage.

6. ACKNOWLEDGEMENT

The project was funded by the US National Science Foundation (NSF) under the contract number NSF of 0131893. We would like to thank the project administrators at the City and County of Denver for granting permission to use their data.

7. REFERENCES

- [1] Amhar, F., J. Josef, and C. Ries, 1998. The generation of true orthophotos using a 3D building model in conjunction with a conventional DTM, *Int. Arch. Photogrammetry. Remote Sensing*, pt. 4, vol. 32, pp. 16–22.
- [2] Biason, A., S. Dequal, and A. Lingua, 2004. A new procedure for the automatic production of true orthophotos, *Int. Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 35, O. Altan, Ed., Jul. 12–23.
- [3] Braun, C., T.H. Kolbe, F. Lang, W. Schickler, V. Steinhage, A.B. Cremers, W. Förstner, and L. Plümer, 1995. Models for photogrammetric building reconstruction. *Computers & Graphics*, vol. 19, no. 1, pp. 109–118.
- [4] Breunig, M., 1996. Integration of Spatial Information for Geo-Information Systems. Berlin, Germany: Springer-Verlag.
- [5] Ermes, P., F.A. van den Heuvel and G. Vosselman, 1999. A Photogrammetric Measurement Method Using CSG Models, *Proceedings of the ISPRS Workshop "Measurements Project Modeling and Documentation in architecture and Industry"*, Vol. XXXII, pp. 36-42.
- [6] Gerke, M., C. Heipke and B. M. Straub, 2001. Building Extraction from Aerial Imagery using a Generic Scene Model and Invariant Geometric Moments. In *Proceedings of the IEEE/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas*, 8-9 November 2001, Rome, Italy, pp. 250-254.
- [7] Graz, M. G., 1999. Managing large 3D urban databases. *Photogrammetric Week*. Stuggart, Germany, pp. 341-349.
- [8] Gruen, A. and X. Wang, 1998. CC-Modeler: A topology generator for 3D city models, *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 53, pp. 286-295.
- [9] Gülch, E., H. Müller, and T. Läbe, 1999. Integration of Automatic Processes Into Semi-Automatic Building Extraction. *Proceedings of ISPRS Conference Automatic Extraction of GIS*

- Objects From Digital Imagery*, September 8 - 10, 1999, Technical University, Muenchen, Germany.
- [10] Lang, F., & W. Förstner, 1996. 3D-city modelling with a digital one eye stereo system, *International Archives of Photogrammetry and Remote Sensing*, 31(Part B3): 415–420.
- [11] Passini, R. and K. Jacobsen, Accuracy analysis of digital orthophotos from very height resolution imagery, *Int. Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 35, O. Altan, Ed., Jul. 12–23, 2004.
- [12] Schickler, W., 1998, Operational Procedure for automatic true Orthophoto Generation. *International Archives of Photogrammetry and Remote Sensing*, 32(4), pp.527-532
- [13] Tseng, Y.-H and S. Wang, 2003. Semi-automated Building Extraction Based on CSG Model-Image Fitting, *Photogrammetric Engineering & Remote Sensing*, vol. 69, no. 2, pp. 171-180.
- [14] Vosselman, G., and H. Veldhuis, 1999. Mapping by dragging and fitting of wire-frame models. *Photogrammetric Engineering & Remote Sensing*, vol. 65, no. 7, pp. 769–776.
- [15] Zhou, G., W. Chen, J. Kelmelis, and D. Zhang; 2005. A Comprehensive Study on Urban True Orthorectification. *IEEE Transactions on Geoscience and Remote sensing*, vol. 43, no. 9, pp. 2138 – 2147
- [16] Zlatanova, S., 2000. 3D GIS for Urban Development. *Dissertation*, Graz, Austria

Table 1: Modification of orientation parameter

	Exterior Orientation Parameter					
	Pose (unit: arc)			Location (unit: feet)		
	φ	ω	κ	X_s	Y_s	Z_s
Initial value	0.0104	0.0158	-1.5503	3143007.320	1696340.371	9032.012
The value derived without accurate improvement	-0.0025	-0.0405	-1.5546	3143041.418	1696562.615	9070.651
The value derived with accurate improvement	-0.0016	-0.0298	-1.5538	3143040.555	1696520.925	9072.272

BIRDS OF A FEATHER

Titles and Authors

Interdisciplinary Analysis of Digital Government Work
Scholl, Hans J (Jochen); Mai, Jens-Erik; Fidel, Raya

E-Government Measurement and Evaluation
Luna-Reyes, Luis Felipe; Gil-Garcia, J. Ramon

XML for Web Site Management in Government: State of the Art and Future Research
Gil-Garcia, J. Ramon; Canestraro, Donna; Costello, Jim; Baker, Andrea; Werthmuller, Derek

e-Governance as a Global Knowledge-Enabling Platform
Finger, Prof. Matthias; Misuraca, Gianluca; Rossel, Dr Pierre

Using System Dynamics for Theory Building in Digital Government Research: Exploring the Dynamics of Digital Government Evolution
Martinez-Moyano, Ignacio J

Citizen Relationship Management: Understanding, Challenges and Impact
Schellong, Alexander

Interdisciplinary Analysis of Digital Government Work

Hans J (Jochen) Scholl, Jens-Erik Mai, and Raya Fidel

The Information School, University of Washington

Box 352840, Seattle, WA 98195-2840, USA

1.206.685.9937

{jscholl; jemai; fidelr}@u.washington.edu

ABSTRACT

This bird-of-a-feather session attempts to break interdisciplinary ground in the context of work content, workflow, and work context analysis in Digital Government. The authors argue that using and connecting multiple theories and disciplines might yield more robust results and deeper understanding of the Digital Government evolution than strictly disciplinary research.

Categories and Subject Descriptors

K.4 {Computers and Society}: K.4.2 Social Issues – K.4.2 Organizational Impacts – K.4.4 Electronic Commerce

D.2.11 {Software Engineering}: - Software Architectures - Domain-specific architectures.

General Terms

Design, Human Factors, Theory

Keywords

Work Content, Workflow, Work Context, Business Process Analysis, Stakeholder Theory, Institutional Theory, Actor-Network Theory, Structuration Theory, Cognitive Work Analysis, Cognitive Systems Engineering, System Dynamics, Soft Systems Methodology, Information Systems.

1. INTRODUCTION

Inter- or multidisciplinary research has been proposed (Hess *et al.*, 2005) when studying “wicked” problems (Pidd, 2004). Many problems, which Digital Government poses, are of that wicked nature. Among others work analysis (WA), that is, the analyses of work content, workflow, and work context have been of special interest in this regard. However, truly inter- or multidisciplinary research designs are generally, but also specifically in Digital Government research in short supply. Such designs would necessitate the active involvement and shared interest of various disciplines when studying any given phenomenon (Hess *et al.*, 2005). Exemplars of interdisciplinary digital government research designs have not yet surfaced. We propose to use a bird-of-a-feather session to discuss what disciplines and what theories would have the capacity to inform WA in digital government. We assume that those theories and disciplines taken together will yield a richer picture and deeper understanding of the phenomenon than when studying the phenomenon with those theories and disciplines in isolation. For the interdisciplinary effort to be successful, though, researchers from the various contributing strands need to bring about an integrated research design, which fits the needs and expectations of the schools and scholars involved. This bird-of-a-feather session attempts to break interdisciplinary ground with respect

to work content, workflow, and work context in Digital Government research.

2. Work Analysis: Content, Flow, and Context

Various theories have been used for studying the dimensions of digital government work. For example, high-level business processes such as public-sector procurement and revenues have been analyzed by means and methods also used in the private sector (Pardo & Scholl, 2002; Pardo *et al.*, 2000). These studies were based on the process analysis and change literatures (Champy, 1995; T. H. Davenport & Short, 1990; Thomas H. Davenport & Stoddard, 1994; Hammer, 1996; Hammer & Champy, 1993) and the work flow analysis and redesign literatures (Cichocki, 1998; Stohr & Zhao, 1999; Zhao *et al.*, 2000). While early business process redesign studies portray the innovative application of information technology as cornerstone of any process and workflow redesign (Fuglseth & Gronhaug, 1997; Grover *et al.*, 1998; Hofacker & Vetschera, 2001; Jarvenpaa & Stoddard, 1998; Kallio *et al.*, 1999), newer studies acknowledge that factors other than information technology may play important and sometimes unpredictable roles in that context (Kumar & Strehlow, 2004; Mansar & Reijers, 2005; Reijers & Liman Mansar, 2005). Common to most traditional WA studies is an emphasis of the functional and technical sides of the problem under study. Factors of cognitive, social, and organizational action, interaction, and constraints remain widely unaccounted for.

3. Related Fields and Theories

More theories and fields than mentioned in this section inform work analysis in digital government: Management Science, Organizational Theory, Sociology, Information Science, and Psychology are among those disciplines, which offer important elements of understanding. Our selection, although incomplete, exemplifies the breadth and depth of theories relevant and informative to the domain of study.

Stakeholder Theory (Donaldson & Preston, 1995; Freeman, 1984; Frooman, 1999; Mitchell *et al.*, 1997) argues that various constituencies influence organizational behavior and success relative to the degree they are able to affect or can be affected by organizational action. Hence, technology-induced change in order to be successful needs to take into account various constituents ‘stakes.’

Institutional Theory (DiMaggio & Powell, 1983; Scott, 1992) holds that institutions are systems composed of cultural, cognitive, normative, and regulative elements, which provide stability and meaning to the resources and activities associated with them. According to Institutional Theory, coercive

(constraining), mimetic (cloning), and normative (learning) mechanisms diffuse over a field of organizations. Digital Government diffusion and its impact on work content, workflow, and work context can be viewed as a case in point.

Actor-Network Theory (ANT), which roots in the works of Callon and Latour (Callon, 1986; Callon & Latour, 1981), provides a lens through which social and technological advances are viewed as coevolving within heterogeneous networks. ANT studies those advances by employing the principles of agnosticism, generalized symmetry, and free association, which provide a conceptual frame for human and non-human ‘actants’ and their interactions without any a priori preference towards the technical or social side (Tatnall & Gilding, 1999)). For work analysis in Digital Government, ANT potentially provides a key to understanding the interplay between human actors and technology.

Structuration Theory (Giddens, 1984) connects human agency and social structure to explain social action. Structuration in Giddens’ words refers to the “the structuring of social relations across time and space, in virtue of the duality of structure” (p. 376). Structure shapes human action, and vice versa. Through structure, human agents interact in habitual, reflexive, reflective, and conscious fashion. Hence, social actors, while constrained by technology, on the one hand, through its use shape that very technology in practice, on the other hand (Orlikowski, 1992). Thus, this duality in nature must also be assumed for digital government and its diffusion.

Cognitive Work Analysis (CWA) is a conceptual framework (Fidel & Pejtersen, 2004; Fidel et al., 2004; Pejtersen, 1985; Pejtersen & Rasmussen, 1997; Rasmussen, 1986; Rasmussen et al., 1994; Vicente, 1999), which helps analyze the work people do, the tasks they perform, the decisions they make, their information behavior, and the context in which they perform their work, for example, for the purpose of information systems design. The CWA framework offers a mechanism for transferring results from an in-depth analysis of human-information-work interaction directly to information system design requirements.

System Dynamics (SD) (Forrester, 1961, 1969a, 1969b, 1975; Richardson, 1991; Sterman, 2000) and also *Soft Systems Methodology (SSM)* (Checkland, 1981; Checkland & Holwell, 1998; Checkland & Scholes, 1990, 1999) provide a systems and feedback-loop perspective for analyzing socio technical phenomena. Technology diffusion and adoption as with Digital Government can be understood in terms of reinforcing or counterbalancing influences between circularly connected variables. While SD analyzes the phenomena by means of formal quantitative modeling and computer simulation, SSM strives to change and improve an organizational situation by a systemic process of sense making, action, and reflection on the action. In digital government both approaches might have the capacity to further and more deeply understanding the work content, workflow, and work context.

As mentioned above, the theories and fields listed here only serve as examples of a wider range of potential theoretical and practical contributions. Furthermore, the explanatory range of contributing theories needs to distinguish the levels of analysis: Those theories may belong to different levels of analysis, that is, to individual, micro, mezzo, or macro levels. It is not well

understood and needs further discussion, in which way, if any, those differences in levels can be overcome.

4. References

- Callon, M. (1986). Some elements of a sociology of translation: Domestication of the scallops and the fishermen of St. Brieuc Bay. In J. Law (Ed.), *Power, action and belief: A new sociology of knowledge?* (pp. 196–233). London: Routledge.
- Callon, M., & Latour, B. (1981). Unscrewing the big leviathan, or how actors macro-structure reality and how sociologists help them to do so. In K. Knorr-Cetina & A. Cicourel (Eds.), *Advances in social theory and methodology: Toward an integration of micro and macro sociologies* (pp. 277–303). London: Routledge and Kegan Paul.
- Champy, J. (1995). *Reengineering management: The mandate for new leadership* (1st ed.). New York: HarperBusiness.
- Checkland, P. (1981). *Systems thinking, systems practice*. Chichester Sussex; New York: J. Wiley.
- Checkland, P., & Holwell, S. (1998). *Information, systems, and information systems: Making sense of the field*. Chichester; New York: Wiley.
- Checkland, P., & Scholes, J. (1990). *Soft systems methodology in action*. Chichester, West Sussex, England; New York: Wiley.
- Checkland, P., & Scholes, J. (1999). *Soft systems methodology in action: A 30-year retrospective* ([New ed.]). Chichester, Eng.; New York: Wiley.
- Cichocki, A. (1998). *Workflow and process automation: Concepts and technology*. Boston: Kluwer Academic Publishers.
- Davenport, T. H., & Short, J. E. (1990). The new industrial reengineering: Information technology and business process redesign. *Sloan Management Review*, 31(Summer), 11-27.
- Davenport, T. H., & Stoddard, D. B. (1994). Reengineering: Business change of mythic proportions? *MIS Quarterly*, 18(2), 121-127.
- DiMaggio, P. J., & Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review*, 48(2), 147-160.
- Donaldson, T., & Preston, L. E. (1995). The stakeholder theory of the corporation: Concepts, evidence, and implications. *Academy of Management Review*, 20(1), 63-91.
- Fidel, R., & Pejtersen, A. M. (2004). From information behaviour research to the design of information systems: The cognitive work analysis framework. *Information Research*, 10(1), paper 210.
- Fidel, R., Pejtersen, A. M., Cleal, B., & Bruce, H. (2004). A multi-dimensional approach to the study of human-information interaction: A case study of collaborative information retrieval. *Journal of the American Society for Information Science and Technology*, 55(11), 939-953.
- Forrester, J. W. (1961). *Industrial dynamics*. Cambridge, Mass.: M.I.T. Press.

- Forrester, J. W. (1969a). *Principles of systems; text and workbook chapters 1 through 10*. Cambridge, Mass.: Wright-Allen Press.
- Forrester, J. W. (1969b). *Urban dynamics*. Cambridge, Mass.: M.I.T. Press.
- Forrester, J. W. (1975). *Collected papers of jay w. Forrester*. Cambridge, Mass.: Wright-Allen Press.
- Freeman, R. E. (1984). *Strategic management: A stakeholder approach*. Boston: Pitman.
- Frooman, J. (1999). Stakeholder influence strategies. *Academy of Management Review*, 24(2), p191 115p.
- Fuglseth, A. M., & Gronhaug, K. (1997). It-enabled redesign of complex and dynamic business processes: The case of bank credit evaluation. *Omega*, 25(1), 93.
- Giddens, A. (1984). *The constitution of society: Outline of the theory of structuration*. Berkeley: University of California Press.
- Grover, V., Teng, J., Segars, A. H., & Fiedler, K. (1998). The influence of information technology diffusion and business process change on perceived productivity: The is executive's perspective. *Information & Management*, 34(3), 141.
- Hammer, M. (1996). *Beyond reengineering: How the process-centered organization is changing our work and our lives* (1st ed.). New York: HarperBusiness.
- Hammer, M., & Champy, J. (1993). *Reengineering the corporation: A manifesto for business revolution* (1st ed.). New York, NY: HarperBusiness.
- Hess, T., Figge, S., Hanekop, H., Hess, T., Hochstatter, I., Hogrefe, D., et al. (2005). Technische Möglichkeiten und Akzeptanz mobiler Anwendungen? Eine interdisziplinäre Betrachtung (technical potential and acceptance of mobile applications? An interdisciplinary perspective-in German). *Wirtschaftsinformatik*, 47(1), 6-16.
- Hofacker, I., & Vetschera, R. (2001). Algorithmical approaches to business process design. *Computers & Operations Research*, 28(13), 1253.
- Jarvenpaa, S. L., & Stoddard, D. B. (1998). Business process redesign: Radical and evolutionary change. *Journal of Business Research*, 41(1), 15.
- Kallio, J., Saarinen, T., Salo, S., Tinnila, M., & Vepsäläinen, A. P. J. (1999). Drivers and tracers of business process changes. *The Journal of Strategic Information Systems*, 8(2), 125.
- Kumar, S., & Strehlow, R. (2004). Business process redesign as a tool for organizational development. *Technovation*, 24(11), 853.
- Mansar, S. L., & Reijers, H. A. (2005). Best practices in business process redesign: Validation of a redesign framework. *Computers in Industry*, 56(5), 457.
- Mitchell, R. K., Agle, B. R., & Wood, D. J. (1997). Toward a theory of stakeholder identification and salience. Defining the principle of who and what really counts. *Academy of Management Review*, 22(4), 853-866.
- Orlikowski, W. J. (1992). The duality of technology: Rethinking the concept of technology in organizations. *Organization Science*, 3(3), 398-427.
- Pardo, T. A., & Scholl, H. J. J. (2002). *Walking atop the cliffs: Avoiding failure and reducing risk in large-scale e-government projects*. Paper presented at the Proceedings on the 35th Hawaiian International Conference on System Sciences, Hawaii.
- Pardo, T. A., Scholl, H. J. J., Cook, M. E., Connelly, D. R., & Dawes, S. S. (2000). *New york state central accounting system stakeholder needs analysis*. Albany, NY: Center for Technology in Government.
- Pejtersen, A. M. (1985). Implications of users' value perception for the design of a bibliographic retrieval system. In J. C. Agrawal & P. Zunde (Eds.), *Empirical foundations of information and software science* (pp. 23-37). New York: Plenum Press.
- Pejtersen, A. M., & Rasmussen, J. (1997). Ecological information systems: Coupling work domain information to user characteristics. In M. Helander, T. K. Landauer & P. V. Prabhu (Eds.), *Handbook of human-computer interaction* (2nd, completely rev. ed., pp. 315-345). Amsterdam; New York: Elsevier.
- Pidd, M. (2004). *Computer simulation in management science* (5th ed.). Hoboken, NJ: Wiley.
- Rasmussen, J. (1986). *Information processing and human-machine interaction: An approach to cognitive engineering*. New York: North-Holland.
- Rasmussen, J., Pejtersen, A. M., & Goodstein, L. P. (1994). *Cognitive systems engineering*. New York: Wiley.
- Reijers, H. A., & Liman Mansar, S. (2005). Best practices in business process redesign: An overview and qualitative evaluation of successful redesign heuristics. *Omega*, 33(4), 283.
- Richardson, G. P. (1991). *Feedback thought in social science and systems theory*. Philadelphia: University of Pennsylvania Press.
- Scott, W. R. (1992). *Organizations: Rational, natural, and open systems* (3rd ed.). Englewood Cliffs, N.J.: Prentice Hall.
- Sterman, J. (2000). *Business dynamics: Systems thinking and modeling for a complex world*. Boston: Irwin/McGraw-Hill.
- Stohr, E. A., & Zhao, J. L. (1999). Temporal workflow management in a claim handling system. *Software Engineering Notes*, 24(2), 187-195.
- Tatnall, A., & Gilding, A. (1999, December, 1 to 3). *Actor-network theory and information systems research*. Paper presented at the Tenth Australasian Conference on Information Systems, Wellington, NZ.
- Vicente, K. J. (1999). *Cognitive work analysis: Toward safe, productive, and healthy computer-based work*. Mahwah, N.J.: Lawrence Erlbaum Associates.
- Zhao, J. L., Kumar, A., & Stohr, E. A. (2000). Workflow-centric communication mechanisms for organizational knowledge distribution. *Journal of Management Information Systems*, 17(3), 45-72.

E-Government Measurement and Evaluation

Luis Felipe Luna-Reyes
Universidad de las Americas
Business School, NE 221J
Santa Catarina Martir
Cholula, Puebla, Mexico 72820
+52 (222) 229-2000 ext. 4536
luisf.luna@udlap.mx

J. Ramon Gil-Garcia
Center for Technology in Government
University at Albany, SUNY
187 Wolf Road, Suite 301
Albany, New York 12205
(518) 442-4473
jgil-garcia@ctg.albany.edu

ABSTRACT

The goal of this ‘Birds of a Feather’ session is to exchange ideas about some important issues associated with the measurement and assessment of electronic government initiatives. The conversation will address topics such as current practices, their strengths and weaknesses, approaches to measure efficiency, effectiveness, quality, and social or economic impacts.

Categories and Subject Descriptors

C. 4 [Computer Systems Organization]: Performance of Systems – *measurement techniques, performance attributes*.

General Terms

Management, Measurement, Performance, Standardization.

Keywords

E-Government assessment and evaluation, Performance indicators, project evaluation, measuring instruments.

1. SESSION DESCRIPTION

Information and Communication Technology (ICT) applications are extending to many aspects of social life. The application of ICT to government, or e-government, has been considered as an important strategy for government reform. E-government has the potential to transform relationships among government, citizens, private organizations, and other stakeholders. However, there is no agreement about a methodology to measure or a framework to assess and compare e-Government development and its impact on social and/or economic development. Current efforts conducted by independent researchers, government agencies or international organizations offer results that are hard to compare given that measurement objectives and indexes respond to a variety of interests and contexts. Moreover, some countries (such as

developing nations) perceive that these international efforts do not properly reflect the level of development of their e-government efforts.

In this way, the purpose of the session will be to analyze the state of the art of e-government measurement and evaluation. It will identify the main dimensions used to measure e-government performance, quality, and social/economic impact, as well as differences, similarities, advantages and disadvantages of the current approaches to measurement. It will also discuss how to use such measurement systems to help public managers and project leaders to improve policy design and decision making about e-government.

The discussion will be guided by questions such as:

- Which are the known experiences in e-government evaluation? Which of them have been more influential? Which of them have been more useful?
- Which are the main similarities between approaches used to compare or rank countries according to their e-government efforts? Which are their main differences?
- Which are the main similarities between approaches used to compare or rank countries and those used to rank states or local governments? Which are the main differences?
- What are the most important dimensions to be included in e-government evaluation efforts? Why?
- How useful are these assessments to public managers and other decision-makers? Is it possible to develop an assessment framework capable to guide policy and decision making? How? What would be its main characteristics?

XML for Web Site Management in Government: State of the Art and Future Research

J. Ramon Gil-Garcia, Donna Canestraro,
Jim Costello, Andrea Baker, and Derek Werthmuller

Center for Technology in Government
University at Albany, SUNY
187 Wolf Road, Suite 301
Albany, New York 12205
(518) 442-4473

jgil-garcia@ctg.albany.edu

ABSTRACT

The goal of this ‘Birds of a Feather’ session will be to share our current knowledge about XML for web site content management in government settings and explore some areas for future research. Specifically, the session will address topics such as the benefits and challenges faced by public managers when attempting to use XML for managing their web sites. The exploration of theoretical frameworks as well as research methodologies will also be considered for the discussion. Finally, the participants will be encouraged to share their insights about future research on this topic.

Categories and Subject Descriptors

H.4.2. [Information Systems Applications]: Types of Systems – *e-government applications*.

General Terms

Management, Performance, Economics, Human Factors, Theory.

Keywords

XML, Content Management, Publication Process, Web Sites, E-Government.

1. Background for the session

Web sites and portals have been an important component in e-government strategies all around the world [4, 5, 6, 7, 12, 13]. As these web sites grow in size and complexity, it is necessary to develop better approaches to web site management and publication processes. HTML, which is the architecture for most existing web sites, was conceived as a way to create single web pages, but presents serious limitations when managing complex

web sites [9]. The future of e-government depends in part on the ability of governments to manage their web sites in a more effective and efficient way. Prohibitive maintenance costs, lack of consistency, and limited capacity to provide multiple formats are just some of the problems that many government web sites are already facing or will face in the near future.

XML (eXtensible Markup Language) is a simplified version of the Standard Generalized Markup Language and it is mainly designed for the Web [2]. XML is generally understood to be a new technology that supports effective data management and exchange between applications [11]. However, XML has another value that is much less exploited or understood – it offers an innovative long-term solution to many of the shortcomings of HTML because it structures and describes web content in a meaningful way [3, 9].

Despite clear advantages, organizations confront many obstacles to the adoption and implementation of XML-based web site management [1, 2]. Similar to other IT initiatives, the factors seem to include technical training and infrastructure readiness, but also solid business case justifications, understanding the impact of organizational change, leadership buy-in, and a firm understanding of where to begin [2, 3, 8, 10].

Thus, this ‘Birds of a Feather’ session will discuss our current knowledge about XML for web site content management in government settings and explore future areas for research. It will describe some of the benefits and challenges faced by public managers when attempting to use XML for managing their web sites. The discussion will also address some theoretical and methodological issues and opportunities. It will be guided by questions such as:

1. How does the use of XML impact the development and management of government web sites? Why?
2. Based on the existing literature, what are the main promises of XML for web site content management in government settings?
3. Are there any other options that provide similar benefits?

4. What are the main challenges or constraints embedded in government settings that may hinder or even neutralize the benefits of XML for web site management?
5. How can return on investment (ROI) for XML be effectively measured and compared across initiatives?
6. What are some useful theoretical frameworks to study the use of XML in government settings? What are the specific advantages and disadvantages of some of the alternatives?
7. What would be effective ways to help public managers to understand the benefits and challenges of XML for web site management?
8. What would be relevant areas of research that need to be explored in order to develop a more comprehensive understanding of this phenomenon?

2. Selected References

- [1] Chen, A. N. K., LaBrie, R. C., and Shao, B. B. M., "An XML Adoption Framework for Electronic Business", *Journal of Electronic Commerce Research*, 4(1), 2003, 1-14.
- [2] Chen, M., "Factors Affecting the Adoption and Diffusion of XML and Web services Standards for E-Business Systems", *International Journal of Human-Computer Studies*, 58, 2003, 259-279.
- [3] Costello, J., Adhya, S., Gil-García, J. R., Pardo, T. A., and Werthmuller, D., *Beyond Data Exchange: XML as a Web Site Workflow and Content Management Technology*. Paper presented at the 2004 Annual Meeting of the Academy of Management: Creating Actionable Knowledge, New Orleans, LA, USA, 2004, August 6-11.
- [4] Detlor, B., and Finn, K., Towards a Framework for Government Portal Design: The Government, Citizen and Portal Perspectives. In Å. Grönlund (Ed.), *Electronic Government: Design, Applications & Management*. Idea Group Publishing, Hershey, PA, 2002.
- [5] Fletcher, P. D., Policy and Portals. In W. J. McIver & A. K. Elmagarmid (Eds.), *Advances in Digital Government: Technology, Human Factors, and Policy*. Kluwer Academic Press, Norwell, MA, 2002.
- [6] Fletcher, P. D., Portals and Policy: Implications of Electronic Access to U.S. Federal Government Information and Services. In A. Pavlichev & G. D. Garson (Eds.), *Digital Government: Principles and Best Practices* (pp. 52-62). Idea Group Publishing, Hershey, PA, 2004.
- [7] Gant, D. B., Gant, J. P., and Johnson, C. L., *State Web Portals: Delivering and Financing E-Service*, The PricewaterhouseCoopers Endowment for The Business of Government, Arlington, VA, 2002.
- [8] Gil-García, J. R., and Pardo, T. A., "E-Government Success Factors: Mapping Practical Tools to Theoretical Foundations", *Government Information Quarterly*, 22(2), 2005, 187-216.
- [9] Hoelzer, S., Schweiger, R. K., Boettcher, H. A., Tafazzoli, A. G., and Dudeck, J., "Value of XML in the Implementation of Clinical Practice Guidelines - The Issue of Content Retrieval and Presentation", *Medical Informatics & The Internet in Medicine*, 26(2), 2001, 131-146.
- [10] Jiang, J., and Klein, G., "Software development risks to project effectiveness", *The Journal of Systems and Software*, 52, 2000, 3-10.
- [11] Kendall, J. E., and Kendall, K. E., "Information Delivery Systems: An Exploration of Web Pull and Push Technologies", *Communications of the AIS*, 1(Article 14), 1999, 1-43.
- [12] OECD, *The e-Government Imperative*, Organisation for Economic Co-operation and Development, Paris, France, 2003.
- [13] Scavo, C., World Wide Web Site Design and Use in Public Management. In G. D. Garson (Ed.), *Public Information Technology: Policy and Management Issues*. Idea Group Publishing, Hershey, PA, 2003.

e-Governance as a Global Knowledge-Enabling Platform

Prof. Matthias Finger
EPFL-CDM-MIR
Odyssea, EPFL Station 5
Lausanne CH 1015
+41 021 6930001

matthias.finger@epfl.ch

Gianluca Misuraca
EPFL-CDM-MIR-e-Gov
Odyssea, EPFL Station 5
Lausanne CH 1015
+41 021 693 0011

gianluca.misuraca@epfl.ch

Dr Pierre Rossel
EPFL-CDM-MIR
Odyssea, EPFL Station 5
Lausanne CH 1015
+41 021 6930002

pierre.rossel@epfl.ch

Building on the experience of the First Edition of the “Executive Master in e-Governance - e-gov”, managed by the Chair of Management of Network Industries ([MIR](#)) of the College of Management of Technology ([CdM](#)) of the [EPFL](#) – Ecole Polytechnique Fédérale de Lausanne (<http://egov.epfl.ch>), the proposal is to discuss the conceptual and theoretical framework developed during the Master and underpinning its second edition, which vision is to create institutional designs through exploring innovative approaches in management and knowledge sharing in public and private organizations.

After discussing the multiple relations among ICTs and governance, the roundtable should focus on the analysis of the use of ICTs in administrations and network industries, and in particular the conceptualization of e-Governance as a growing phenomenon that is emerging within public and private sector institutions around the world and as a significant discipline within the field of public management and public-private partnership-building.

In order to better understand e-Governance (as a concept and in practice) the debates regarding this issue, that are most often polarized between those who feel that ICTs will enhance the participation of citizens in the government policy decision-making process, and those who feel that it will simply be “business as usual” via a new medium should be analysed. Starting from the argument that e-Governance will enhance the economic performance of countries, and especially developing ones, either through the creation/strengthening of ICT industry, or through the easier access of private sector to crucial public sector information, the debate about e-Governance should instead go beyond the consideration of simply increasing administration performance and the capacity for public service delivery, but focusing more on the “networked” relations among the different stakeholders involved in managing ICT-related affairs.

In this regard, the discussion should build upon a review of current literature that reveals how e-Governance can be explored from several perspectives, thus highlighting the multidisciplinary nature of it.

Within this context, it is proposed to present the innovative conceptual framework upon which the EPFL Executive Master in e-Governance is based. In particular, its objective of establishing a global network of excellence in research and training in e-Governance, built on a number of international key institutions but “open” to potentially all research and training institutions in this field and connected to international partners, eager to provide Executive Professionals with a robust knowledge-base and practical skills to face the challenging nature of managing ICTs in a changing world.

The “knowledge-enabling platform” on which the “e-gov” Master is based upon is therefore:

- multi-dimensional: investigating the political, economic and social impacts of ICTs;
- multi-sectoral: including several areas of industry and the public sector;
- multi-stakeholder: involving academia, civil society, governments and private sector organizations;
- multi-disciplinary: investigating the relations between the development and integration of ICTs in government and changes in patterns of social organization and cultural orientation,
- multi-level: focusing on the different levels of governance, from local to supra-national, identifying the implications of ICTs for national, regional and local development; and
- multi-functional: analysing the broader range of governance-related issues, from service-delivery to regulation and policy-making.

Using System Dynamics for Theory Building in Digital Government Research: Exploring the Dynamics of Digital Government Evolution

Ignacio J. Martinez-Moyano

Argonne National Laboratory
9700 S. Cass Avenue, Building 900
Argonne, IL 60439
630.252.8824

imartinez@anl.gov

ABSTRACT

This theory-building effort is designed to attract people with an interest in digital government evolution by using formal modeling in the social sciences. By using the system dynamics approach in a group setting, the main drivers of digital government organizational and technological sophistication are explored and formally articulated in a causal structure. The results are embodied in a formal model that can be mathematically characterized to conduct numerical simulations.

Categories and Subject Descriptors

J.4 [Social and Behavioral Sciences]

General Terms

Experimentation

Keywords

System dynamics, e-government, systems thinking, formal modeling

1. INTRODUCTION

This theory-building effort is designed to capture underlying structures that drive digital government evolution by looking at the interacting nature of factors that drive digital government organizational and technological sophistication. This effort builds on previous work that studied e-government evolution as a function of the interaction of rules, norms, and individual preferences for action [3] and attempts to expand and refine the basic dynamic hypothesis posited there.

The theory-building effort is conducted by using the simplified version of group model building scripts proposed by Andersen and Richardson [1]. Discussing the drivers of digital government with a group of researchers interested in this topic will make it

possible to blend different perspectives on the subject and create a shared view of the drivers of these dynamics. This mechanism is a way to create operationally the concept of shared vision as proposed by Senge [7] and to expand knowledge in this area beyond what individuals could achieve alone.

2. APPROACH

The system dynamics approach can be used to study the evolution of digital government because it provides a methodology for exploring feedback-rich systems in which the nature of the relationships among the elements creates circular causality. System dynamics provides a framework for investigating the effect of changes in one variable on other variables over time. System dynamics, a computer-aided approach to policy analysis and design, applies to dynamic problems arising in complex social, managerial, economic, or ecological systems [5] as is the case in digital government.

3. METHOD

The proposed method is system dynamics modeling in a simplified version of its group model building variant. Group model building deals with “the processes and techniques designed to handle the tangle of problems that arise in trying to involve a large number of people in model construction.” [4: 375] Research that deals with the specifics of this method has been termed ‘group model building’ [1, 2, 6].

The literature in group model building continues to grow and gives specific guidelines and scripts for carrying out these processes [1, 2, 8]. According to Zagone [9], two main types of models may arise in group model building: micro-world and boundary-object models. Micro-world models try to capture objective reality, whereas boundary-object models are artifacts that help in understanding diverse views about the world. In these models, a composite view of reality emerges as a result of the interaction of the individuals that build it.

4. MODEL

A preliminary dynamic hypothesis is provided (see Figure 1) to start the group session and to engage participants in the discussion. This hypothesis acts as a concept model (as proposed by Andersen and Richardson) that will guide the group session.

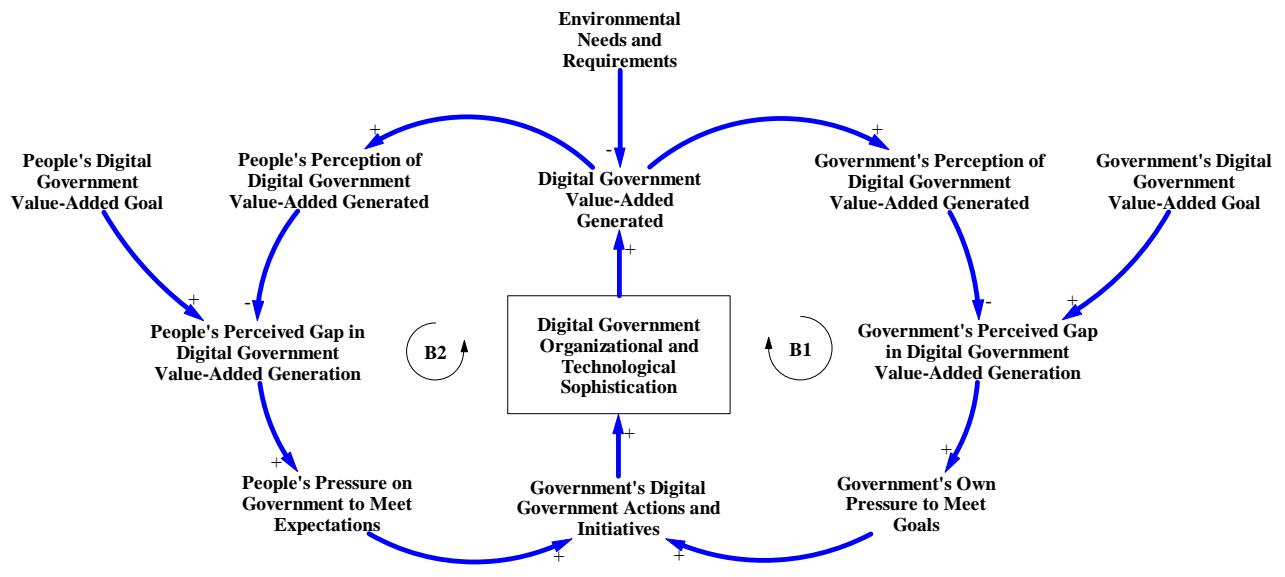


Figure 1. Dynamics of Digital Government Evolution.

In Figure 1, two main feedback mechanisms compete to determine the dynamics of digital government organizational and technological sophistication. One (B1 in Figure 1) is driven by government's recognition of a gap between what is achieved and the goal established. This gap generates pressure for new initiatives, leading to change in digital government. A second mechanism (B2 in Figure 1) is based on the people's perception of the value added by digital government and a comparison of that with expectations set for it. The identified gap is translated into pressure (from the people) to the government to increase value in their digital government endeavors. The two mechanisms, institutional and political, interact and function at the same time, creating the necessary forces for other mechanisms to emerge and for the dynamics of digital government to exist.

5. CONCLUSION

Digital government research is inter- and multi-disciplinary in nature. Bridging the gap that exists between practice-based digital government research and the creation of a sound theoretical basis for the field is a complex endeavor. The use of system dynamics modeling, focused on the intersection of social, institutional, and behavioral elements, to explore digital government evolution can help bridge such gap.

6. ACKNOWLEDGMENT

Work supported by the U.S. Department of Homeland Security.

7. REFERENCES

- [1] Andersen, D.F. and Richardson, G.P. Scripts for Group Model Building. *System Dynamics Review*, 13 2. (1997), 107-129.
- [2] Andersen, D.F., Richardson, G.P. and Vennix, J.A.M. Group Model Building: Adding More Science to the Craft. *System Dynamics Review*, 13 2. (1997), 187-201.
- [3] Martinez-Moyano, I.J. and Gil-Garcia, J.R. Rules, Norms, and Individual Preferences for Action: An Institutional Framework to Understand the Dynamics of e-Government Evolution. *Lecture Notes in Computer Science*, 3183. (2004), 194-199.
- [4] Richardson, G.P. Citation for winner of the 1999 Jay W Forrester Award: Jac Vennix. *System Dynamics Review*, 15 4. (1999), 375-377.
- [5] Richardson, G.P. System Dynamics. in Gass, S.I. and Harris, C.M. eds. *Encyclopedia of Operations Research and Management Science*, Kluwer Academic Publishers, Boston, MA, 1996, 656-660.
- [6] Richardson, G.P. and Andersen, D.F. Teamwork in group model building. *System Dynamics Review*, 11 2. (1995), 113-137.
- [7] Senge, P.M. *The Fifth Discipline: the Art and Practice of the Learning Organization*. Doubleday/Currency, New York, 1990.
- [8] Vennix, J.A.M. Group Model-Building: Tackling Messy Problems. *System Dynamics Review*, 15 4. (1999), 379-401.
- [9] Zagonel, A.S., Model Conceptualization in Group Model Building: A Review of the Literature Exploring the Tension Between Representing Reality and Negotiating a Social Order. in *Proceedings of the 20th International Conference of the System Dynamics Society*, (Palermo, Italy, 2002).

Citizen Relationship Management: Understanding, Challenges and Impact

Alexander Schellong
Harvard University

Johann Wolfgang Goethe- University, Frankfurt a. M.
alexander_schellong@ksg.harvard.edu
schellong@em.uni-frankfurt.de

ABSTRACT

Customer Relationship Management shares the same objective of improving citizen orientation in public administration with digital government and new public management. The latter concepts are criticized for not reaching their desired objective and focusing on the inside instead of integrating citizens. CiRM can finally add a clear customer strategy to government. Citizen Relationship Management (CiRM) refers to a cluster of management practices, channel and IT solutions based on CRM found in Marketing. Goals can be improving citizen orientation, better accountability and changing the citizen government relationship. CiRM lacks a theoretical and conceptual basis due to sparse empirical and theoretical academic research. At the moment CRM's technological side is mainly applied in combination with 311 call centers, its impact yet unknown.

Categories and Subject Descriptors

General Terms

CRM, Citizen Relationship Management, Cross-boundary collaboration, multi-jurisdictional, Change Management, Accountability, Business process reengineering, privacy, performance management,

Keywords

Public services, 311, call center, portal, CiRM, CRM, multi-channel, networked governance

1. INTRODUCTION

Citizen Relationship Management (CiRM) refers to a cluster of management practices, channel and technological solutions that apply private sector Customer Relationship Management (CRM) in the public sector. Goals can be improving citizen orientation, customer service, increasing accountability and changing the citizen government relationship by embedding citizens opinions throughout public administration instead within a certain processes/ agency, hierarchy level or elected officials. Basic principles of CRM are

personalization (products, information, services), integration (planning processes, business process reengineering, product development), interaction (channels, communication, outreach, surveys), and selection/ segmentation (heavy data analysis, identify the top 20% of customers who make 80% of the profit, termination of unprofitable clients). Moreover, quality/ performance measurements, change management and a strategy promoting customer oriented culture are vital to any CRM concept or project. CiRM lacks theoretical and conceptual clarity due to sparse administrative implementation and empirical or theoretical academic research.

Reviewing existing public sector CiRM projects, CiRM it is now mostly applied in conjunction with 311 call centers or citizen service centers with a focus on its technological component and the municipal level. Interestingly, CRM software has been used in areas such as public owned water and sewer or power supply for some years. Furthermore, agencies offered different ways of interaction such as the counter, call center and more recently the web. However, despite the rise of a customer driven government in the late 80's there is mostly no common customer service strategy, existence, gathering and use of data or attempt to build a closed loop environment. Even eGovernment initiatives fall short of their expectations in this regard.

Thus, there are many questions which need to further discussion with regard to understanding, implementation and impact:

How could we define Citizen Relationship Management? Which aspects of CRM are and which are not applicable in public administration? What are benefits and goals? What are differences and similarities between eGovernment and other customer service initiatives? Is CiRM capable of finally creating a more citizen centered and focused government? Can CiRM change relationships in the triangle citizens, elected officials and administrators? Will citizen be more passive in a democratic sense if they are more satisfied with their day to day government interactions? Will CiRM and 311 projects be sustainable if leadership changes? Is CiRM just a hyped term which will soon be replaced by something else?

FIELD TRIPS

Locations

The Synthesis Center

U.S. Border Patrol Command and Control Intelligence Coordination Center

The Synthesis Center

*A Joint Project between San Diego Supercomputer Center
and The California Institute For Telecommunications and Technology*

Wednesday 5/24/06

Time: 1:00 - 5:30 pm (PDT)

FIELD TRIP ABSTRACT:

In many areas, science is becoming a team sport. Important science questions can only be addressed by bringing together multi-disciplinary groups of scientists along with IT experts. "Big science" in the future will not just involve running individual "super" computations. It will also be about "super" science, which brings together scientists from across sub-disciplines and disciplines, along with heterogeneous databases and tools, to solve multidisciplinary and multi-scale science problems in a collaborative way. This is the notion of synthesis.

Synthesis is at the heart of many complex science endeavors, including numerous NSF-funded projects that are either led by the San Diego Supercomputer Center (SDSC) at the University of California, San Diego (UCSD), or where SDSC has significant involvement. It is at the heart of important efforts funded by NIH and the Department of Homeland Security. The Synthesis Center serves as a center to facilitate interactions and sharing of ideas among scientists on complex science and engineering issues, without relying on a central computing resource. This is the vision for the next generation of science.

What is the Synthesis Center?

The Synthesis Center is a place where groups of collaborating scientists and engineers come together for face-to-

face sessions to directly address science questions using cyberinfrastructure tools. Sessions may run for a day, a few days, or even a few weeks. Even with the availability of a variety of remote collaboration tools and technologies, real science still happens mainly in face-to-face meetings — which the Synthesis Center facilitates. Cyberinfrastructure plays a supporting role, as a facilitator of science.

The **Synthesis Center** is a portal to:

- Cyberinfrastructure and a unique environment consisting of large-scale, wall-sized displays linked to powerful on-demand cluster computing systems, with easy access to storage and important databases;
- Data analysis and mining tools, which may be stored locally or available via high speed networks.

The **Synthesis Center** allows groups of researchers and scientists to meet together in a resource-rich environment where they can address cutting edge science questions that can be solved only by teams of inter-disciplinary scientists meeting together to enable face-to-face discussions and "what if" computations.

The **Synthesis Center** can be scheduled for several days or weeks at a time for

scientists to meet and discuss science in real time with access to on-demand computing. Between such sessions, Center staff work with domain scientists and computer scientists to assemble the data, tools, and systems needed in preparation for the next session projects.

Some projects in which the Synthesis Center are involved:

- High Density Display
- MBT Notebook Project
- SDSC Active Poster
- Cell Signaling Pathways Visualization
- Medical Data Visualization On Cell Phones and PDA's
- Pocket Fold
- Virtual Tour Guide Project

Using the Synthesis Center

There are three ways to use the Synthesis Center:

- Preparing to run experiments. The Synthesis Center can be scheduled for activities related to examining, understanding, and "cleaning" databases in preparation for use in science endeavors. There may also be collaborative efforts to arrive at common metadata specification, and define shared knowledge structures. The Center provides easy access to data and a variety of query and information visualization tools that allow a group of scientists to look at the data and information on wall-sized displays. These sessions can also be used for tool tuning—i.e., to test drive tools, understand their capabilities, and test their correctness prior to their use in serious science and/or large runs.

- Running experiments. Once databases have been assembled (or connected to real-time streams) and tools have been readied, the Synthesis Center can be used to perform science runs while the collaborative group is assembled at the Center. These science sessions can run for days, even weeks.
- Studying experimental results. For problems that require large computations such as a large simulation run, a science group may first define a cyberinfrastructure campaign, where they design and develop a series of runs in advance. The runs may produce terabytes of output, which can be stored in digital libraries. After the campaign completes, the science group assembles in the Synthesis Center to examine the results, visualizing simulation output easily accessible and on an on-demand basis. Results from the campaign can be stored or deleted.

U.S. Border Patrol Command and Control Intelligence Coordination Center

Wednesday 5/24/06

Time: 1:00 - 5:30 pm (PDT)

FIELD TRIP ABSTRACT:

The Command and Control Intelligence Coordination Center (CCICC) was created in support of field operations and decision-makers. The CCICC is a multifaceted intelligence operation encompassing the collection, analysis, dissemination, and command apparatus for several law enforcement entities along the southwest border. The CCICC conducts 24/7 Intelligence operations in support of the Customs and Border Protection (CBP) Mission. This center also contains the capabilities to become the centralized Command and Control facility for the CBP staff during critical events. The CCICC is the only program throughout the US Border Patrol to utilize a spatial decision support

system to implement coordinated enforcement operations. Along with being pioneers in the realm of enforcement technology, the CCICC is also on the leading edge of innovative research, planning, and implementation of the National Incident Management System (NIMS). CCICC is the future of safe, effective, and efficient emergency operations management for natural and man-made disasters. Due to a dynamic all hazards and all disciplines approach, CCICC can rapidly coordinate multiple incident command posts, multiple agencies, multiple emergency operations centers, and maintain the continuity of operations plan for all of San Diego Sector operations.

ABOUT THE DIGITAL GOVERNMENT RESEARCH CENTER



The Digital Government Research Center (DGRC) was established in 1999 with the support of the National Science Foundation. DGRC is a collaboration between the University of Southern California's Information Sciences Institute (USC/ISI) and Columbia University's Computer Science Department. DGRC is headquartered at USC/ISI.

People

A strong team of university-based scientists and developers at DGRC represents expertise in text processing, database information integration, human-computer interaction, knowledge representation, data mining, computer networking, security, ontologies, and other areas relevant to government. Since 1999, projects with collaborators in various other institutions have worked with experts from several Federal government agencies on a variety of research topics.

Center Activities

The center engages in four types of activity:

- Information Technology research: Developing advanced information systems to address critical areas of need for government agencies and citizens in data management and online transactions;
- Digital Government community building: Helping to organize the annual dg.o conferences that bring together staff from federal, state, and local government, researchers in IT and social sciences, and companies with a commercial interest in Digital Government;
- Production of the monthly newsmagazine dgOnline, and coordinating the activities of other major areas of DG research;
- Digital Government program growth: Organizing and participating in workshops to help develop new directions for NSF's Digital Government program.

Research Focus

DGRC research focuses on dealing with large, possibly distributed, sets of numerical and textual data, collected by government agencies at all levels. Processing, inspecting, and integrating information across different sources can be extremely difficult.

DGRC has been investigating several foundational questions in information processing for government use. One is single-point integrated access to large, dispersed collections of government data. Another is in-depth sophisticated analysis of texts and comments sent by the public to the government. A third is semi-automatic assistance to government staff to create and extend their metadata, taxonomies, and glossaries.

Current major undertakings are:

The **Argos** project, with the California Department of Transportation (Caltrans), the Los Angeles Metropolitan Transport Authority (MTA), and others, is integrating information systems under a web services paradigm to help improve analysis and freight flow control in the Los Angeles region.

The **eRule** project, a partnership with researchers at Carnegie Mellon University, University of Pittsburgh, and San Francisco University, is working with the Federal EPA and Department of Transportation to analyze automatically text comments by the public about proposed

regulations. This includes near-duplicate removal, opinion identification, stakeholder identification, and more.

The **SifT/Guspin** project is using machine-learning techniques to automatically map and integrate air quality data collected by various environmental protection agencies in California and elsewhere.

The **OntoGrow** project is working with the US Environmental Protection Agency and the Department of Defense to develop techniques to extract and formalize pertinent aspects of textual descriptions of data, in order to automate linking of text documents and data collections across agencies and countries.

The **QUALEG** project facilitates electronic interactions between citizens and city government. Partners in this project include the Technion in Israel and the city of Saarbrücken in Germany in the context of a larger project, funded by the European Commission's eGovernment unit.

Education, Community Building and Outreach

In addition to its research activities, DGRC is involved in important educational and outreach efforts related to digital government, including:

- DigitalGovernment.org, an online resource for the digital government research community and other interested parties. Among other things, it houses a searchable index of NSF Digital Government related awards.
www.digitalgovernment.org
- The annual International Conference on Digital Government Research
www.dgrc.org/dgo2006/
- DG Online, the monthly newsletter of Digital Government Research
www.digitalgovernment.org/news/stories/dgonline_latest.jsp
- International outreach and collaboration. DGRC has played a central role in establishing contacts and collaborations among international researchers in digital government/e-government. DGRC has helped organize several workshops and other meetings for this purpose.

Contacts

Director:	Yigal Arens (arens@isi.edu)
Director for Research:	Eduard Hovy (hovy@isi.edu)
Assistant Director for Development:	Valerie Gregg (vgregg@isi.edu)
Co-Director:	David Waltz (waltz@cs.columbia.edu)
Publicity Officer:	Chrystol Koempel (koempel@isi.edu)

Digital Government Research Center
USC Information Sciences Institute
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292-6695
USA

<http://dgrc.org>

