

# Birds of a Feather Sessions

Organizers: Carol Hert, Syracuse University and Judith Klavans, Columbia University

The Birds of a Feather sessions provided a way for attendees to coalesce into focus groups, voice their opinions on current and future directions, and incorporate involvement from related communities to extend the impact of the program. The intent is to identify those issues that the community as a whole should address, and if possible, to suggest strategies to move forward. Each group provided a summary of its discussion.

## **Group One: Collaboration in the Digital Government Program: How can different communities work together towards a common goal?**

**Participants:** Sharon Dawes (chair), Jack Marshall, Lung Teng Hu, Jim Thompson, Joe Danek, Yigal Arens, Teresa Harrison, Sibel Adali, Bob Maslyn, Norman Dowe, Sal Stolfe, Vanghn Blankship, David Cheney, Alan MacEachren, Peter Bloniarz, Melvyn Ciment, H. Michael Chung.

### **Discussion**

Collaborations are usually difficult to achieve, for reasons ranging from social to technical. Despite the need for Government agency workers and researchers to communicate, transfer technical results and software, etc., there is currently no easy way of doing so: data security, firewalls, and limitations of collaboration software are but a few of the inhibiting factors. Interested parties discuss how to overcome the obstacles and achieve effective collaborations using new IT, e.g. with common workspaces. Our group consisted of three people in government, seven people in the research field, and two people from the private sector.

We discussed many aspects of collaboration including: researchers in the same discipline, researchers in different disciplines, and the third issue was research among government agencies. We felt that collaboration on tools and services needs to be done to address these issues. Research on collaboration was essential, and an ongoing community of interest was important. In future sessions we would want to address more case studies, more on focused discussions, as well as a faculty development session on building more research collaborations.

## **Group Two: Citizens and Users and Societal Impacts**

**Participants:** Stuart Shulman (chair), Bob Carlitz, Fred Conrad, Roslin Hauck, Eric Welch

### **Discussion**

This group examined research issues related to the provision of services and information to users and the potential societal impacts of digital government. It might include issues of user tasks, user knowledge, user expectations, social impacts of technology, etc. Interface and desired functionalities of systems might also be relevant. The focus is on users that are not performing highly expert statistical tasks. There were three top issues that are most pressing:

1. Access and Usability. Access and usability are significant problems for citizens. For example, at BLS, stats are available, but users do not access them (they give up for various reasons). Access and usability are affected by IT literacy of citizens and bureaucrats.
2. Education and Innovation. Rulemaking online requires expansion of how organization presents the process. We Need to rethink the education process for public activities that are moved online. Do agency staff like to open up the process? What are the contributors / barriers to greater openness?

3. Organizational Change and community development. To what extent does enhanced ability of communication between government and citizen affect the ability of government to function effectively and efficiently? How does digital government affect the legitimacy of citizen comments, legitimacy of what government does in response, and the legitimacy of online data?

#### Solutions:

1. Usability engineering. More usability testing is required. Usability engineering must be an important element of design. More collaboration is needed among stakeholders such that developers understand not only the users, but also the structure of the data and the work processes applied. Ways in which agencies can overcome silo effects and isolation must be developed.
2. Including educational components in other programs. Lifelong learning and retooling to help agency personnel responsible for developing the activities for such activities as rulemaking is required.
3. Longitudinal studies of organizational and community change as interactivities of internet will help understand user needs.
4. Digital signatures. Different types of security technologies.

### **Group Three: Presentation and Display for Universal Usability**

**Participants:** Judith Klavans (chair), Lois DelCombre, John Looney, Gary Marchionini, John Pearson, Neil Scott.

#### **Discussion**

This group explored research issues related to the presentation of complex information from agencies. We covered universal usability concerns, technical underpinnings of presentation, presentation formats and their use, provision of help, etc. Of particular concern is Section 508, accessibility for people with disabilities.

This group talked about making presentations that make statistics readable, even making them designed by users. We also discussed the presentation of information to those individuals who are sight impaired. Issues of how to jump start research to help agencies produce a product they want were covered.

We covered the fact that the digital government program provides an important platform for the development of new presentation technologies, but that the limitations of agency involvement greatly inhibit our testing of these technologies. We would like to see more agency activity, in terms of funding and access to information so that we have a better idea of what is needed. We would like the program to explore ways to encourage technology transfer using programmers, rather than graduate students. The academic structure is not set up to build bullet-proof systems, although it is our goal and wish to design new ways to meet user needs. This is the domain of commercialization, which is the link missing from the current program. Finally, an important concern of this group was the way information could be misused, particularly for Section 508. What ways are being developed to protect users from misuse of their access information?

### **Group Four: Digital Government Technologies: What's Next?**

**Participants:** Leana Golubchik (chair), Jamie Callan (scribe), Vijay Atlira, Jamie Callan, Catherine Dippo, Ali Farahani, Leana Golubchik, Kincho Law, Scott Midkiff, Charlie Rothwell

#### **Discussion**

The digital government program currently funds the development of technologies such as data integration, heterogeneous database access, diverse data types, security and confidentiality both in data gathering and dissemination, etc. Yet at the same time, there are a range of related technologies that could be included. This group considered which technologies need to be further developed, better exploited and, which research domains related to these technologies should be included.

**Technology Transfer.** Technology transfer activities related to Digital Government projects are badly underfunded. Current funding is for feasibility studies and prototypes only; some government partners understand this, some do not. Graduate students are the wrong resource for building application systems. Agencies should fund technology transfers, but may not have the budget or appropriate technical or other resources. Agencies would also rather buy than develop. Current research funding does not go far enough to demonstrate commercial viability so that companies would pick up research, or so that VCs or others would fund startups. Need something like SBIRs more tightly integrated with Digital Government program. The Digital Government program needs to get a slice of the \$200 million allocated for cross-agency projects; Digital Government projects have already been peer-reviewed and evaluated in operational environments, so some sort of “fast track” status is appropriate. Projects need greater visibility beyond current DigiGov partners, to entice other agencies, or so that multiple agencies could share development costs.

**Business Models (Business Process Re-engineering).** Even when the technology path is clear, potential new business models (i.e., new ways of doing business) for government agencies are unclear. Who is the “customer” for a system that monitors people on parole? It is the agencies that need more help developing business models they are completely overlooked, and underfunded. One example of this is: putting probation officers in schools (because that’s where many of the offenders are)—is this effective? Or is it a useful or counterproductive resource for schools, how does it interact with other programs? The group feels that there should be an emphasis on pilot programs that would be widely applicable and that could be copied easily if successful (e.g., at county level). We compare this to monitoring non-violent drug offenders on probation; monitoring via postcards is ineffective (low response rate); new program is kiosk related, so participation is more easily enforced; could this be ported to the Internet. Does it make sense? Digital divide issues (does this community have appropriate access). What would be better business models for this population? One idea is that we need more Business School types to study reengineering of government services.

**More cross-jurisdictional portals.** There is a need for more topic-specific state-wide information portals, e.g., transportation. This is partly a technology issue, partly a matter of integrating a breadth of resources. User-specific focus, as opposed to agency-specific. Might tie into business process reengineering projects. There are all of these kinds of projects that need to be done

**Privacy.** We felt that if there were unique identifiers that it would be very useful. There are however, major privacy issues, such as authentication being a key issue; anonymity; many policy issues, not really the domain of CS folks; the main issue is social costs; the more interrelated systems become, the more serious are the consequences of hackers, identity-theft, etc; there is a need for people to study the ripple effects of errors, hacks, etc, in inter-related systems, i.e., how errors propagate, and how errors can be corrected

**Large-scale integration of heterogeneous resources.** Current projects rely heavily on unified schemas, global ontologies, and manual wrapping of databases. We feel that these projects will become the Digital Government success stories because the advantage would be to have a one-stop shopping and integration of services. The disadvantage of course is that these approaches don’t scale (economically) to large numbers of agencies, because manually wrapping hundreds or thousands of resources is not practical (expensive). Also, integrated systems are even more difficult to change in the future, because a local change ripples into other systems. The success stories will be difficult to reproduce on a large scale. We need more emphasis on automatic wrapping of resources and automatic use of ontologies, as USC/ISI project is doing; this is a hard problem, but really the only approach that is viable in the long-term and on a large scale.

**Scalable infrastructure.** Government agencies need to shift to sharing hardware/network resources, similar to Web-hosting services, so that excess capacity is shared, and expertise is localized.

**Digital divide.** There were five main categories that we felt would put people when dealing with the Digital Divide: income, disabilities, rural areas, and, user issues, e.g., mental health issues (not everyone is easily trainable), and age. What is the technology that needs to be “basic service” for all people? The size of the digital divide depends upon the services that we’re talking about: email, Web, Word, Access, Java, etc, a person will all have different levels; what is required for what levels of participation? We are aware that some digital divide issues not necessarily difficult to attack. Solving some (possibly easy) digital divide issues may help in

attacking other “social divides” that have been difficult to make progress on in the past. Attacking the digital divide in these cases is a surrogate for solving other problems.

**Personalization.** In order to address this issue we need more emphasis on mass customization of data, so that the “right” information is delivered to people without cookies (privacy issues). One approach would be to deliver a large chunk of data to the client it may require bandwidth, do customization at the client (more private). Another approach would be encrypted queries (client reveals information in a way that the server can process without understanding), encrypted data (server returns data that a client can process without having full access to the database).

**Multilingual access.** One major problem with the digital divide is due to the multilingual problems (not everyone speaks English). We need to make sure to have Enabling technology that has the capability to have a Cross-Lingual Information Retrieval (CLIR) i.e., Spanish queries return English documents; Automatic machine translation (MT), because it is impractical (too expensive, too slow) to translate every document into every language. The Goal is to have automatic translation for greater dissemination, but government requires great accuracy; what happens when people make decisions based on incorrect translations? How are the choices of no information (because a document isn’t available in a target language) balanced against available, but possibly inaccurate information (due to translation errors)? We feel that research on government-specific issues in MT and CLIR is needed. User interface issues (how are limits of translation accuracy conveyed, how is the possibility of multiple, somewhat different translations conveyed)?

**Information distillation.** In today’s day and age it is now so easy to get more information than a person can (or wants) to read. Many people give up before finding what they need and people need software that will help them analyze or distill out the information that they need. We need enabling technology: information filtering at the client level; basic research is fairly far along, but little research so far on deploying in personal systems: question answering software at the server; this is an emerging research area, capable of extracting simple, fact-based answers from retrieved text documents; more research is needed to handle the kinds of questions people would ask of government portals. clustering of search results (to show what’s out there) and iterative retrieval (to drill down along a particular dimension). Drill-down interfaces: How do we make them easy for people to use and understand?

**Automatic analysis of public input.** How does an agency handle a million user responses to a request for public feedback? The answer: text categorization to sort people into bins (by viewpoints or other characteristics); effect analysis to identify level of emotion; and, personalization of (email) responses from government officials

**Policy consequences of increased public participation.** This area seems as if it’s under studied. There should be more rapid public feedback that may imply less time for deliberation. There should also be more opportunity for demagogues. As well as authentication: Is a person providing feedback from your city, state, country? How does authentication balance against anonymity (although historically the government ignores anonymous feedback)

## **Group Five: Ontologies: Building Useful Knowledge Bases across Agencies**

**Participants:** Eduard Hovy (chair), Shawn Bowers (scribe), Bob MacGregor, Raphael Malyankar, Shawn Kerrigan, Melania Degeratu, Jose Luis Ambite, Steve Reichenbach, Isabelle Cruz, Bill Labiosa.

### **Discussion**

The digital government program includes research on technologies to automatically build and populate large ontologies. These conceptual knowledge bases can be used by people to navigate complex terminology both within and across agencies. The goal of this group was to identify both theoretical and practical problems in building ontologies, propose concrete solutions, and explore additional data resources that can be analyzed for ontology construction. We also considered issues of integration with metadata, both existing and proposed. What follows is our discussion.

It seems that the uses of ontologies people are working with include: Data in heterogeneous DB where values need to be integrated. For example, what is a price? Nomenclature, glossaries, and terms ie, do terms mean the same thing? We need some kind of definitional standardization (similar to machine translation of language), and to do reasoning, knowledge representation, infer relationships in knowledge, etc. There are different definitions of ontologies. We do not deal with the pure philosophy notion of ontologies. There is danger of philosophical wars about what they should be.

Our top three issues are as follows:

**1. Problems:** Different words that mean the same thing (e.g., different spellings, or synonyms), but, different words with the same meanings (acronym); acronym lists, hierarchies specific to a discipline; science and engineering are different things; pushing to end user to decide context is important.

Co-occurrence analysis, statistical analysis (computational linguistics) patterns of co-occurrence. of words, bank tellers, money, etc; bank, bait, fish, water, etc.; co-occurrence - new research in ontology's.

**2. Has anyone actually built ontologies from scratch using text?** There has been some intervention, but in general this can be easier than from text mixtures of domain experts and software engineer's. There is no methodology. Are there tools to support ontologies for creating, browsing, APIs, etc? There are some standards being developed for notation (e.g., KIF) some experience, browsing some standards for specific domains on standardizing ontologies. Several standards and mapping between them is a potentially better approach than coming up with "the standard ontology"

**3. Building ontologies** The data exists, but there is research being done to build applications that manipulate it an the emphasis on concepts is usually wrong. We can't come up with a single definition for attributes that applies for all cases. Specifying each different definition typically leads to an incomplete list. When people typically model the data (e.g., define a schema), they don't necessarily know the exact definitions anyway. Instead the focus should be on the (binary) relationships. We need to look for usage patterns in the relationships, but don't constrain what you can have. One question to consider is; aren't you just moving the problem to relationships? The answer is no, because we aren't interested in the semantics of concepts, just observe and correlate the relationships. There was a discussion about knowing monthly and yearly salaries. Basically, the discussion was about what the difference is between Entity-Relationship Diagrams/UML Diagrams and ontologies. Parallel ontologies: Are there (declarative) languages to express mapping between parallel ontologies, to traverse back and forth.

We felt it useful to have some survey if it wasn't too difficult to list projects that were being done in government. From simple vocabularies to more complex structures. For example, who created the NAICS ontology of industries? This would be one place to start. We need a list of ontology efforts in government plus contact points. What should have been addressed? We need more on the possibility/trade-offs of creating ontologies programmatically as opposed to by hand? What has been tried and what has failed? What should be the scope of an ontology? An argument can be made that the ontology only needs to be as complex as the problem you need to solve, but you can create large backdrop ontologies, where you can plug in your localized ones (big, general high-level tools)

## **Group Six: Metadata: Current Issues in Building Successful Metadata**

**Participants:** Carol Hert (chair), David Barker-Plummer, Tony Stefanidis, Mathew Weaver, Athman Bouguettaya,

The success of digital government initiatives often depends on the ability to combine multiple data sources, share knowledge about those data sources and enable access to them. Metadata are critical for the successful accomplishment of these tasks. Current issues in metadata systems such as appropriate metadata standards and structures, location of metadata, and evaluation of metadata repositories will be discussed.

## Discussion

The session started with introductions expressing our diverse interests in metadata and metadata usage and then, on to a discussion of the definition of metadata, of which we all seemed to have our own. We had differences in what would be considered metadata, how to model relationships among metadata, data, and other information.

Particular points of interest during this discussion included:

- Rationales and their capture as metadata
- The semantics of metadata within a database and across databases
- Metadata definition must be within a context. Different uses for metadata (for search and retrieval, for rationale recording and use, for integration of databases, etc.)

Agencies have a different idea of metadata. For example, how to do cognitive tests and the rationale behind survey questions, etc. The derived data—called metadata by survey folks at workshop. Statistical people “This column has a variance of X” —think this is metadata, but we do not think that it is. Should there be a distinction between derived and metadata? What if you do an analysis on data? For any survey instrument, there are levels of data. Derived data is below the collected data. Metadata is above the collected data. Sometimes there is meta-metadata above the metadata.

There are two problems: semantics within the database, and then semantics across databases. The schema is the metadata, so schema integration data is meta-metadata. This problem really has been plaguing us. We have used an ontological approach, but we are not satisfied. The semantics that you chose may not be the semantics that were intended. Agencies that collect data is geared from the producer perspective rather than the user perspective. They sometimes neglect the metadata in process design. Users could be very diverse—a citizen planning a vacation, someone trying to respond to a natural disaster, a real estate agent building a virtual reality of a neighborhood. The question is, are we dealing with structured or unstructured data? The problems are different for each. In structured data (i.e., databases) the schema is already there. In the case of semi-structured data, you still have to build the schema. People feel more comfortable with schemas because you can reason about schemas, query schemas, etc. We are dealing mostly with databases. The problem of metadata itself is not such a big problem—except for understanding the schema. We want to have a uniform interface to interact with the semi-structured and the structured data. Our main problem is to automatically look at the schema and generate the metadata (ontologies).

## **Group Seven: Geospatial Data: A special case**

**Leader:** David Chase (chair), Ken Lanfear, USGS, Ilya Zaslavsky, San Diego Super Computer Center, Richard Muntz, UCLA, Sherrie Harms, Univ. of Missouri-Columbia, Hanan Samet, Univ. of Maryland-College Park, Francisco Artigas, Rutgers Univ., Peggy Agouris, University of Maine-Orono, Soon Chun, Rutgers Univ., Katherine Hansen, Council on Excellence in Government, Panayotis Partsiuvelos, Univ. of Maine, Ron Li, Ohio State Univ, Chaitan Baru, San Diego Supercomputer Center, Brad Parks, Univ. of Colorado.

This group considered the research issues in the utilization and dissemination of the increasingly large amount of geospatial data. How can this community gain better access to existing sources, provide guidance in its manipulation and dissemination to users of all expertise.

## Discussion

While the group did not come to any consensus as to the most important issues in the area of GIS, there were many issues presented for consideration. These can be categorized in three general areas: Data Content; Data Access; and Data Context.

Data Content Issues were: Accuracy; matching data sources; Resolution; Metadata (temporal as well as spatial content); Remote sensing

Data Access issues were: Standards; Infrastructure; Real-time data; Section 508, non-visual presentation; Providing wider access while maintaining expert knowledge Over-dependence on off-the-shelf software

Data Context issues were; selection among datasets (appropriate to user context); static vs. dynamic (adaptable to changing context); context-sensitive resolution; integration with non-spatial datasets and applications; decision-making tools.

While no specific next steps emerged from the discussion, the group agreed to the value in maintaining a dialog among the group members. An e-mail list has been created by David Chase for the group. You may email David if you wish to be added to this discussion list (David\_E.\_Chase@hud.gov).