Toward a Statistical Knowledge Network

Gary Marchionini, Stephanie Haas, Catherine Plaisant, Ben Shneiderman, & Carol Hert
[march, haas]@ils.unc.edu, [plaisant, ben]@cs.umd.edu, cahert@syr.edu

- Type: Full Paper
- Demo as well?:   Yes, there will be a series of demos for the various interfaces
- Author(s): see above
- Address: 100 Manning Hall, UNC-Chapel Hill 27599

Statistics support planning and decision making and enormous efforts are made to collect data and produce statistics at all levels of governance. An important principle of democratic societies is that government statistics should be accessible to the broadest possible constituencies to empower better plans and decisions in all aspects of life. Given the potential of near-ubiquitous Internet access in homes and workplaces and efforts by government agencies to mount websites, physical access to large volumes of government is close to a fait accompli. What remains a significant challenge is enabling access to the right statistical information at the right time and in the right form. This challenge has several facets with accompanying implications:

- There is a massive volume of federal, state, and local statistical information; finding the right data imposes a requirement for good filtering support.
- Federal, state, and local statistical providers use a variety of standards, formats, terminology, and practices for collecting and disseminating statistics; making these data appear seamless requires metadata interoperation.
- The full population includes people with diverse user needs, experiences, and technological platforms; serving the full range of people from novice to expert raises the need for multiple, alternative solutions.
- The level of statistical/numerical literacy in the population is generally low; helping people to find and understand statistical information implies providing online help and support.

These challenges will be met over time by education and better systems. One approach to incorporate a learning population with improving technology is to create a statistical knowledge network (SKN). There is substantial literature on knowledge management (e.g., Nonaka, 1991; Davenport & Prusak, 1997; McInerney, 2002) that speaks to the need to integrate information processes together with information products. A statistical knowledge network must likewise incorporate the people and processes that gather, analyze, manage, and report statistical information into an accessible framework. From the end user's perspective, such a network should also be concept-based rather than file-based.

Our approach is to envision user interfaces as the glue in the SKN. Our view of user interfaces is not simply as front ends to client-side services pasted on in ad hoc fashion, but rather as integrally coupled data and agile user-controllable views and mechanisms. Our design is driven by several principles (Shneiderman, 1998):

- Practice user-centered design driven by needs analysis (know the user)
- Provide alternative views and control mechanisms (support universal access)
- Strive for highly interactive systems that allow penalty-free explanations and look aheads (create direct manipulation, dynamic query interfaces)
- Map specialized vocabularies to end-user vocabularies
- Provide on-demand help and support
- Usability test designs at all stages of development.

Our aim is to empower people to "find what they need and understand what they find." Such designs present several challenges. Our experience with highly interactive interfaces in a variety of settings

(Shneiderman, 1997; Plaisant et al. 1997a, 1997b; Greene et al., 2000; Marchionini et al., 2000) demonstrate the urgency of good, consistent metadata that can be transferred to the client for easy manipulation. A related requirement is to develop useful partitions of the data to help users visualize/recognize what is and is not available. With these principles and requirements as motivation, we have initiated five threads of work outlined below: user needs scenarios, metadata development, information architecture, vocabulary and concept maps, and online help. These underlying research threads provide the infrastructure for novel interface designs. This paper first presents an architectural overview of our vision of a SKN and then outlines progress on these research threads and preliminary user interfaces that take advantage of this work.

SKN Architecture

In the current situation, most statistical data providers have sophisticated systems that may include large scale commercial (e.g., SPSS, SAS) or customized software, database and file management software, and various LAN and Internet server software. Typically, these data systems are behind firewalls to preserve confidentiality and provide security. Some agencies may have integrated metadata, some have separate metadata files distinct from the primary data files, and still others may participate in shared metadata repositories. Within agencies and across agencies there are often membranes of varying permeability; some departments or agencies share much data and systems behind the main firewall and others share little or none. These systems are expensive legacy systems that will continue to be operational for the foreseeable future. Each agency may create a publicly available website that provides various access mechanisms (e.g., text search, navigable link structures) that return webpages (e.g., files with reports, tables, etc.). In the current model, each end user interacts with the agency website or in the case of FedStats, an intermediary website that provides access to the webpages in the participating agency sites.

Just as web services have necessarily become highly layered to allow specialized functionality and management to be encapsulated at local levels while insuring that basic communication can take place across disparate systems, we propose a layered architecture SKN that depends upon intermediary services to link back end statistical systems with end users. This architecture is depicted in Figure 1. In this architecture, a public intermediary integrates human resources (domain experts and end user communities that may answer questions or contribute new resources and data) with ontologies for statistical concepts
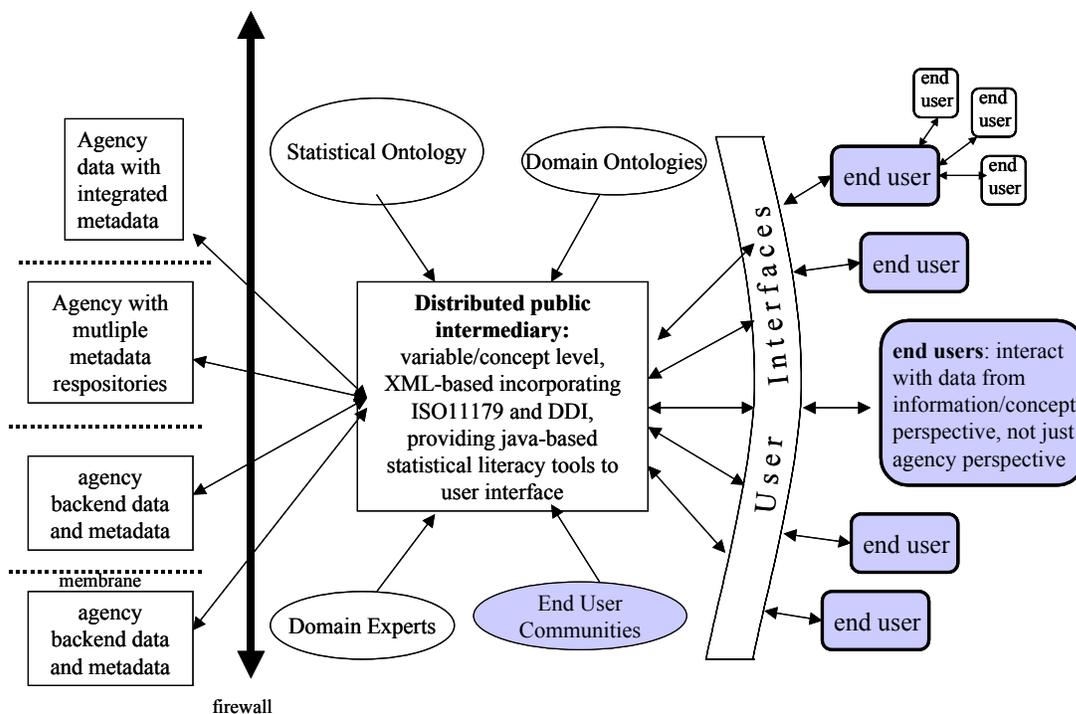
and various domains (e.g., health, economics). The public intermediary is envisioned as an XML-based mapping of these resources onto cleansed data and metadata made available by participating agencies. The public intermediary supports a variety of end-user interfaces for exploring, searching, manipulating, and studying statistical concepts and associated data. Individual end users or end-user communities, or value added services use the interfaces of their choice to access all types of statistical information rather than specific interfaces for each statistical agency and data set. To date, we have made progress on each of the research threads that underlie the public intermediary and user interfaces.

Scenarios and User Needs.
Twenty end user scenarios were developed early in the project to identify key issues in integration and serve as design guides for preliminary work. We aimed to identify realistic problem-based scenarios that would require statistical data from multiple agencies, across multiple levels of government. We conducted brainstorming sessions (via email) with our agency partners to identify interesting and typical situations. Our past work with statistical agencies informed this work and twenty candidate scenarios were identified. These were refined to fifteen after preliminary discussions and searches. A template to collect data was developed and a team of researchers then conducted extensive searches in the WWW and in various commercial indexes to find pertinent data for 'answering' five of the scenarios.   The details of these efforts are given in a working paper available at http://ils.unc.edu/govstat/papers/scenario_paper_nov_14_2002.doc. We used the results to query agency partners about what data they could provide that would meet the needs expressed in the scenarios. These scenarios demonstrated the many facets of questions and possibly pertinent information, the range of public and private resources available, and the difficulties associated with vocabulary, quality assurance, and acquiring the actual data from various sources. In addition to the scenario development, a user study of the FedStats website was conducted to test how well people are able to understand vocabulary in its index (Ceapura, 2003).

Metadata
Metadata sits at the heart of the SKN. We have been investigating metadata options available in the statistical community (e.g., the Data Documentation Initiative and the ISO 11179 standard), what metadata exists in our partner agency systems, and have conducted a user study on what metadata experts and novices find useful for different statistical problems. See http://ils.unc.edu/~ohjs/stats.html for a primer on statistical metadata standards and sources, Hert and Haas (this conference) for results from the metadata study and http://ils.unc.edu/govstat/papers/hert-statistics.ppt for a recent presentation. Our goal is twofold: to leverage metadata to build easy to use search and browse services for the SKN; and to provide explanatory information at the survey, variable, and statistic levels to help people understand the data they find.

Information Architecture
Alternative ways to slice and dice large data sets address user diversity as well as helping all individuals to understand the overall structure of the data. Organizing and naming data and their anchors are the key challenges of Information Architecture (IA) and a classical problem of indexing. A research team conducted an intensive investigation of the Energy Information Administration (EIA) website with the aims of becoming intimate with the details of tens of thousands of webpages and creating a concept map of what is included. The resulting three tiered organization (Fry & Su, 2003) will serve as an alternative view to the current EIA website and also be the basis for instantiating one of the interface prototypes (Relation Browser).

Hand-crafting an organization and set of identifiers for a large website (let alone for the aggregation of all these sites that would feed the SKN), is extremely expensive and difficult to scale. We have been investigating different ways to automatically categorize all the objects in a website. Our first approach was to crawl an agency website (BLS), take the 100K+ unique strings and after a set of reductions (locate

words using WordNet, stem, apply stop list), compute term frequency (tf), term-document frequency (tdf), and term-frequency/inverse document frequency (TFIDF) for each term. The first 100 principal components of the terms were projected onto the TFIDF term-document matrix, and this result was subjected to kmeans clustering with and without the 100 most commonly occurring terms in the collection. The resulting clusters yielded promising 'slices' through the corpus and we are experimenting with labeling techniques combining manual and automatic (e.g., centroid based) methods (Efron, Zhang, & Marchionini, in review). The goal is to automatically create an organization and labeling scheme that can be compared to manual and agency-created schemes. The comparisons will be made using the Relation Browser and other UI prototypes.

Vocabulary
We have begun developing a statistical ontology to relate statistical concepts. See http://ils.unc.edu/govstat/papers/Santa-Fe012303.ppt for an overview of the plan and status. A first step is to create a statistical glossary and begin to develop graphical explanations (including flash animations to explain concepts such as 'seasonally adjusted.' The ontology supports the creation of explanations and will allow users to explore statistical concepts and the relationships between them. Classes of terms include: statistical concepts, date/time, geography, topics, and user terms. We are investigating the following types of representations for definitions: examples, brief tutorials, demonstrations, interactive simulations, pointers to related concepts, and live links to community or agency personnel (Haas et al., in review).

Help
The WWW environment is mainly self serve—online help has been largely ignored. We aim to address this by developing help and support services for the SKN. One approach is to use context-dependent sticky note help (see Plaisant et al., 2003) to provide on-demand help for users. We are investigating animated demos (see Dominick et al., 2003) as a technique as well. The animated glossary help noted above is another approach. Finally, we will investigate ways to link metadata (e.g., source, units of measure, expanded explanation) to data values to help people understand what they have found.

User Interface Prototypes
The research threads above are meant to support highly interactive interfaces. We have initiated several prototype designs. Figures 1-6 depict our preliminary designs. Figure 1 shows one version of the Relation Browser (Marchionini & Brunk, in press) that allows users to mouse over attributes in a column of values and see how many objects exist for each category (slice) through the database on another set of attributes. Clicking on one of the bars returns the objects. Figure 2 shows a more advanced version that allows users to explore topical as well as numeric and geographical attribute sets through mouse over mechanisms. Figure 3 shows a geographic map that displays results for a variety of attributes set through sliders and other control mechanisms (Golub & Shneiderman, 2002). Figure 4 shows a network view of the relationships in a data set. Users can change the level of detail rotate or zoom to investigate relationships in the webpages or data objects. Figure 5 shows an example of dynamic, context specific sticky note help (Plaisant et al., 2003). Demonstrations for these interfaces will be made at the conference.

Summary and Current Status
In addition to the different work summarized above, we aim to foster relationships with federal and state agencies. One immediate goal is to continue to acquire data sets and the appropriate metadata that can be incorporated in the interface prototypes. Some current issues we aim to address in the coming year include: What metadata are most crucial for end user access and understanding? How can metadata be integrated across agencies without waiting for all to adopt a common standard? Which metadata standards will be adopted and how should these be mapped to interface features? As we scale automatic classification to multiple agency data sets, should the data be merged before clustering or can we cluster on each agency and then merge results (much easier computationally)? What multi-layered help

strategies are most effective? What data objects will be the appropriate and practical units of analysis for design (files, variables, concepts)? As the user interface prototypes are populated with data sets, we will conduct user studies to assess their effectiveness and guide subsequent designs.

A SKN is an ambitious undertaking. Our partner agencies (BLS, Census, EIA, NASS, NCHS, and SSA) have shown great willingness to collaborate toward this goal and we have begun discussions with some state-level agencies. The FedStats consortium has demonstrated that the federal statistical community has the will and skill to work together to better serve people's statistical information needs. Their willingness to work toward a SKN is the logical extension to the success of FedStats and promises to serve even wider portions of the population in the years ahead.

References
1. Ceaparu, I. (2003). Governmental Statistical Data on the Web: A Case Study of FedStats. *IT& Society*, 3(1), 1-17. http://itandsociety.org
2. Davenport, T. & Prusak, L. (1997). Information ecology, NY: Oxford University Press.
3. Dominick, J. Hughes, A. Marchionini, G. Shearer, T. Su, C. and Zhang, J. *Portal Help: Helping People Help Themselves Through Animated Demos*. February 2003. UNC-CH Technical Report SILS TR-2003-01.
4. Efron, M., Zhang, J. & Marchionini, G, (in review) Comparing feature selection criteria for term clustering applications. Toronto, July 28-31, (2003). poster Submitted to "*Proceedings of ACM SIGIR 2003*"
5. Fry, J. & Su, C. (2003). Methodological challenges in information architecture: Adventures in re-indexing federal statistical websites to enhance access to end users. Poster presented at the *Fourth Information Architecture Summit*. Portland, OR (March 2003).
6. Greene, S., Marchionini, G., Plaisant, C., & Shneiderman, B. (2000). Previews and overviews in digital libraries: Designing surrogates to support visual information seeking. Journal of the American Society for Information Science, 51(4), 380-393.
7. Golub, E. & Shneiderman, B. (2002). Dynamic query visualizations on world wide web clients: A DHTML solution for maps and scattergrams. Technical Report: HCIL-2002-08 , CS-TR-4367 , UMIACS-TR-2002-47. http://www.cs.umd.edu/local-cgi-bin/hcil/sr.pl?date=golub
8. Haas, S., Patteulli, C., & Brown, R. (in review). Understanding Statistical Concepts and Terms in Context: The GovStat Ontology and the Statistical Interactive Glossary. Submitted to the *ASIST 2003 Annual Conference*.
9. Marchionini, G. & Brunk, B. (in press). Toward a General Relation Browser: A GUI for Information Architects. *Journal of Digital Information*.
10. Marchionini, G., Brunk, B., Komlodi, A., Conrad, F., & Bosley, J. (2000). Look Before You Click: A Relation Browser for Federal Statistics Websites. *Proceedings of the Annual Meeting of the American Society for Information Science* (Chicago,, Nov. 12-16, 2000), 392-402.
11. McInerney, C. (2002). Knowledge management and the dynamic nature of knowledge. *Journal of the American Society for Information Science*, 53(12), 1009-1018.
12. Nonaka, I. (1991). The knowledge creating company. *Harvard Business Review*, 79(1), 96-104.
13. Plaisant, C., Kang, H., Shneiderman, B., Helping users get started with visual interfaces: multi-layered interfaces, integrated initial guidance and video demonstrations, to appear in *Proceedings of 10th International Conference on Human-Computer Interaction*, Crete, Greece, 22-27 June 2003.
14. Plaisant, C., Marchionini, G., Bruns, T., Komlodi, A., & Campbell, L. (1997a). Bringing treasures to the surface: Iterative design for the Library of Congress National Digital Library Program. *ACM CHI '97 Conference*. (Atlanta, March 22-27, 1997), p. 518-525.
15. Plaisant, C., Shneiderman, B., and Muhslin, R. (1997b) An Information Architecture to Support the Visualization of Personal Histories**.** *Information Processing & Management*, 34, 5, pp. 581-597, 1998.
16. Shneiderman, B., *Designing the User Interface: Strategies for Effective Human-Computer Interaction: Third Edition*, Addison-Wesley Publ. Co., Reading, MA (1998).
17. Shneiderman, B. (1997). Direct Manipulation for Comprehensible, Predictable, and Controllable User Interfaces. Proceedings of IUI97, 1997 *International Conference on Intelligent User Interfaces*, Orlando, FL, January 6-9, 1997, 33-39.
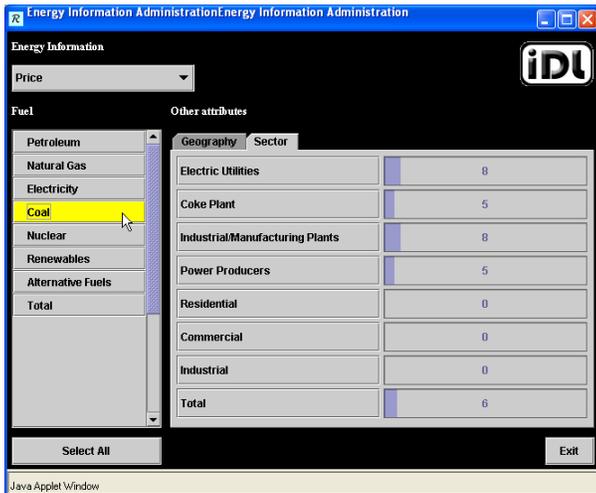
Figure 1. Relation Browser A
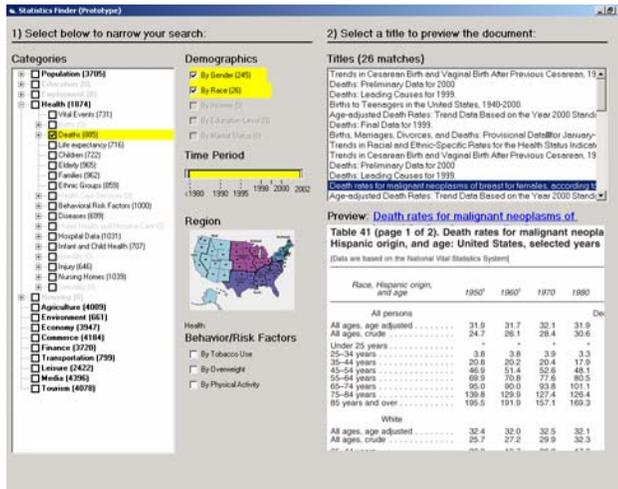

Figure 2. Visual Information Browser A
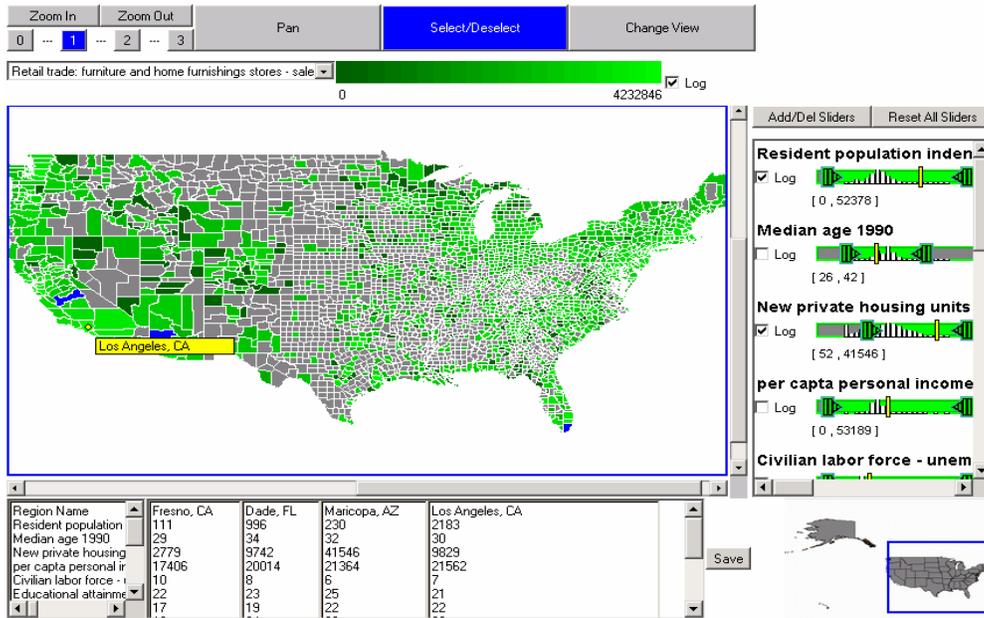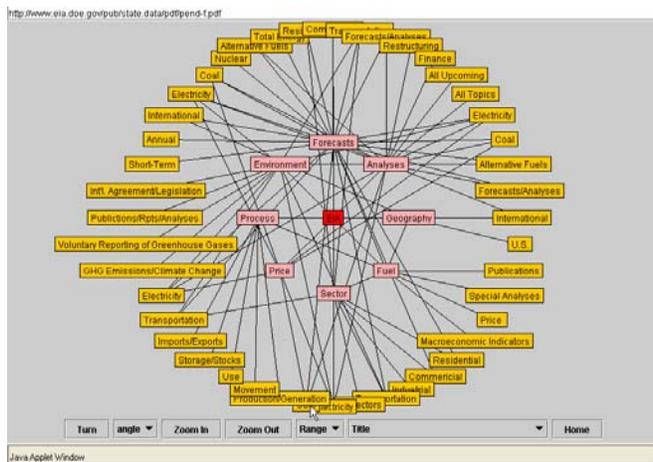

Figure 3.Interactive Map

Figure 4. TableHunter

Figure 5. Sticky-Note Help