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