

# Project Highlight: Toward a Statistical Knowledge Network--2004-05

Gary Marchionini  
Stephanie Haas  
University of North Carolina  
100 Manning  
Chapel Hill, NC 27516  
011.919.966.3611

{march; haas}@ils.unc.edu

Ben Shneiderman  
Catherine Plaisant  
University of Maryland  
HCIL  
College Park, MD  
001.301.405.2680

{ben; plaisant}@cs.umd.edu

Carol Hert  
University of Washington  
Tacoma, WA

011.253-692-5874

cahert@u.washington.edu

## ABSTRACT

This paper summarizes progress made in the third year of a digital government project that aims to develop user interface models and prototypes to help people find and understand government statistics. We envision a Statistical Knowledge Network consisting of repositories of data and tools to support online community access and interaction.

Categories and Subject Descriptors

H.5 [[INFORMATION INTERFACES AND PRESENTATION](#)]

## General Terms

Experimentation, Human Factors, Standardization

## Keywords

User interfaces, digital government, statistical information, information retrieval, online help

## 1. INTRODUCTION

This project aims to help people find and understand government statistical information. To achieve this goal, we envision a statistical knowledge network that brings stakeholders from government at all levels together with citizens who provide or seek statistical information. The linchpin of this network is a series of human-computer interfaces that facilitate information seeking, understanding, and use. In turn, these interfaces depend on high-quality metadata and intra-agency cooperation. In this briefing, we summarize our accomplishments in the third year of the project.

In the second year, a statistical knowledge network architecture was developed in partnership with our statistical agency partners. This year, we aimed to help agencies think about how they can adopt and adapt the model. Ben Shneiderman worked with the National

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.

Conference '04, Month 1-2, 2004, City, State, Country.

Copyright 2004 ACM 1-58113-000-0/00/0004...\$5.00.

Infrastructure for Community Statistics (NICS) group <http://www.brook.edu/metro/umi/federalinformation.htm> to map this architecture to the models of data sharing and integration emerging from their work. This relationship is ongoing. Our other major efforts in the final year of the project fall into three related threads of work: advancing our efforts to prototype, test, and refine user interfaces to support finding and using statistical information (UI thread); creating metadata models and techniques for populating those models with topical descriptors (metadata thread); developing models and interfaces for online help for people using government websites (help thread).

## 2. USER INTERFACE THREAD

One major accomplishment this year was to install a full-featured user interface on a federal website. In cooperation with the FedStats team (Marshall DeBerry and Rachael Laporte Taylor) two instances of the Relation Browser interface were installed on the FedStats server. These instances provide access to the FedStats and BLS websites (see [http://trinity.fedstats.gov/RB/rb\\_fedstats.html](http://trinity.fedstats.gov/RB/rb_fedstats.html) and [http://trinity.fedstats.gov/RB/rb\\_bls.html](http://trinity.fedstats.gov/RB/rb_bls.html) respectively). Based on this installation and subsequent testing, we are currently developing instances for all our other agency partners. Another accomplishment on the user interface thread was continued development of a theoretical design model for sonification and a prototype interface that applies the model to an audio interface for maps. Several papers have been presented and the UI is undergoing user testing (Zhao et al). We received a small supplemental award from NSF to continue this work for 2005-06. Work also continues on developing a general model for user controlled search results categorization interfaces and is the basis for a dissertation (Kules). UI prototypes for an animated statistical glossary (see <http://ils.unc.edu/govstat/papers/glossarylink.html>) were developed and two user studies were conducted, leading to papers (Haas et al.) and a Masters Paper (Wilbur). We are working with Ann Aiken at CDC on implementing some of the animated glossary entries at the NCHS website. Work also continued on developing a design framework for recorded demonstrations (show me demonstrations,

Plaisant et al.). Overall, our UI efforts are slowly finding their way into the practices of government agencies, albeit less readily than we might hope.

### 3. METADATA THREAD

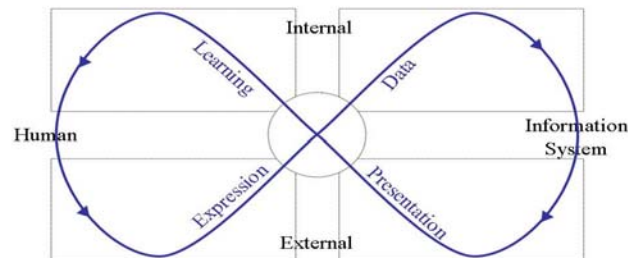
The Metadata thread has several components that have yielded promising results. In the first years of the project, we aimed to develop a DTD based on the DDI specifications that could be easily adopted by agencies and applied to all aspects of statistical data dissemination. The draft DTD was used to mark up several tables and reports, however, the experience of doing the markup demonstrated the enormous effort that will be required within agencies to document at this level. This year, the team worked on a layered model that would allow agencies better control over the level of metadata to add. Several papers and presentations were made in conjunction with Dan Gillman from BLS (Denn et al.). Another component of this thread aims to define a statistical ontology that will drive the metadata practices. Several papers and posters were submitted to demonstrate feasibility (Pattuelli et al.), including ways to link the ontology to project interfaces. The third component of this thread aims to automatically generate topical and geographic categories for statistical websites using machine learning techniques. A Text Mining Toolkit was created and made available as open source code (<http://ils.unc.edu/govstat/demos.html>). The toolkit is under implementation consideration at Census and BLS. The basic process is to crawl the website, apply unsupervised clustering to establish partitions of the site, manually label the partitions and tune the algorithm parameters to create a statistical model that is then used to classify each webpage into one or more of the partitions. The toolkit has been revised over several iterations with BLS, Census, and FedStats data and we are in the process of applying the toolkit to SSA, NASS, NSF, NCHS, and EIA data. Once the classifications are done, the data is piped into a Relation Browser instance.

### 4. HELP THREAD

The help thread has yielded both theoretical and practical results this year. The animated glossary prototypes were user tested and our overarching layered model for online help extended. Likewise, experience with the show me demonstration prototypes have informed this multi-layered model of help. In January, we held a two-day symposium devoted to online help for public service websites (<http://ils.unc.edu/govstat/helpsymposium.html>). The symposium brought attendees from several government agencies as well as academic researchers together to react to our progress to date and to define a research agenda. The symposium and subsequent discussions have yielded a

Human-Information Ecology Model that represents the flow of information between people and information systems (See Figure 1). This model will continue to evolve as we develop user interface instances and study how people use help in agency websites.

**Figure 1. Human-Information Ecology Model**



### 5. CONCLUSION

These three interrelated threads of work aim to make government statistics easier to find and understand. Although we have had good success in engaging our agency partners over the life of the project, a substantial challenge to our technology transfer aims is the attrition of agency partners through retirement. Because collaboration depends on trusting relationships that must be built over time, progress on long-term projects such as this one is more reliable when work involves a combination of senior and junior professionals in agencies. However, we see our success well beyond the direct impact on the agencies we work with; our published research influences a broad range of government and other professionals.

Overall, the third year of this project has shown good maturation as we worked with our agency partners to bring prototypes and theories to practice in the agencies. This has led to good testing and revision and we plan to continue to evaluate these threads in the coming year as the project draws to a conclusion. See <http://www.ils.unc.edu/govstat> for the project website and access to the papers referenced here.

### 6. 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.