

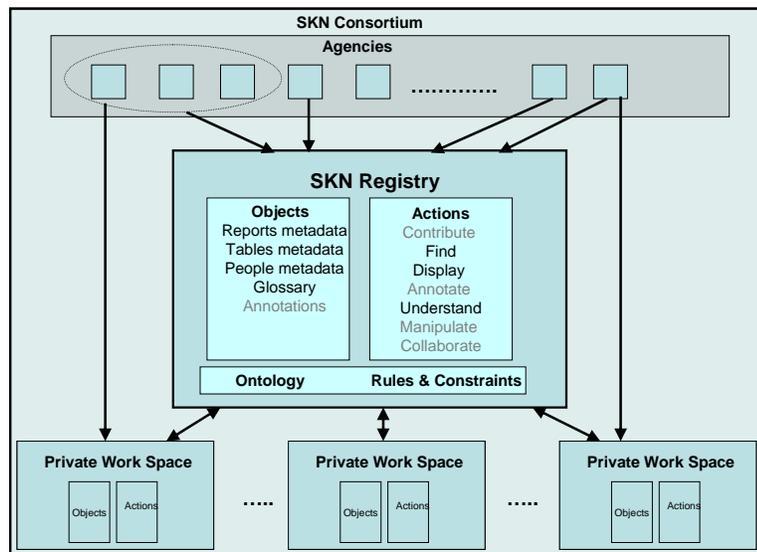
## Project Highlight: Toward the Statistical Knowledge Network

Gary Marchionini University of North Carolina- Chapel Hill march@ils.unc.edu	Stephanie Haas University of North Carolina- Chapel Hill haas@ils.unc.edu	Ben Shneiderman University of Maryland-College Park ben@cs.umd.edu	Catherine Plaisant University of Maryland-College Park plaisant@cs.umd.edu	Carol A. Hert Syracuse University cahert@syr.edu
--	---	--	--	---

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 second year of the project.

Based on meetings with our government partners (BLS, Census, EIA, NASS, NCHS, and SSA), we have defined a consortium and registry-based architecture for the SKN and have made progress on several fronts: interface development and testing, metadata definitions, and automatic classification of statistical resources. The architecture is depicted below.

Figure 1. Statistical Knowledge Network Architecture



The SKN supports a **consortium** of people, organizations, and resources devoted to government statistical information. Individuals may have various roles in the consortium ranging from occasional participants who participate anonymously to those who participate regularly and publicly. Agencies and organizations may likewise participate as occasional contributors or harvesters or may accept formal responsibility for operating the tangible components of the SKN and coordinating its governance. Individuals interact with the SKN from their own **private work spaces** with clients that are able to harvest and use registry objects and tools. The consortium depends on the SKN **registry** of statistical information **objects**, a suite of **actions**, an **ontology**, a set of **rules and constraints**, and associated tools for working with this information and other entities in the consortium.

Any **object** that supports the SKN's purpose may be registered. Objects may be submitted for registration by agencies and individuals, or may be added to the registry as the result of web crawler actions. Example objects include: reports, tables, pointers to people, glossary entries, and annotations. Information on objects will be stored using standardized Extensible Markup Language (XML) structures tied to a Document Type Definition (DTD). It is important to note that in most cases, the registered object will not be actually stored in the registry, but rather in the provider's server. In some cases, such as user annotations or contributions from individuals or organizations

without reliable delivery capabilities, the primary object along with the usual metadata pointers may be stored in the registry in non-government web spaces.

**Actions** are enabled by various tools in the registry. At this time, a series of tools to support common actions is defined. Actions in the registry may be initiated and controlled by users or they may be automated processes that maintain and expand the registry. The following actions and supporting tools are defined: a) **Contribute**. Tools for input of object descriptions using the SKN DTD and for crawling web resources. Includes management actions such as input validation, indexing, record editing and deletion. Supports crawling partners on schedule, version control, security and authentication services. b) **Find**. Tools for searching and browsing/exploring the indexed registry, including actions for sorting and relationship browsing. Includes search engine logs and analysis tools and tools to fetch the objects from resource holders, including possible authentication and commercial transactions. c) **Display**. Tools that act on outputs of the find tools. User-selected objects are retrieved using locator information from the registry indexes and are displayed in a variety of formats (e.g., PDA, full-screen, visual/audio, English/Spanish). Note that find and display may be tightly integrated. Results may be piped or exported in various formats. d) **Annotate**. Tools to enable users to attach comments (and metadata for those comments) to existing objects in the Registry. e) **Understand**. Tools to define and explain objects and actions. Using information on the context of user actions (such as the current active module), deliver context-sensitive definitions, explanations, instructions, and so on to the interface. These tools are always active during a user session with the SKN and may include glossary, context, and user profile management. f) **Manipulate**. Tools to act on outputs of the find and understand modules. Objects can be extracted, compared, visualized, used in statistical analyses and components of objects can be manipulated (e.g., cutting and pasting of columns in a table). g) **Collaborate**. Tools to support users working together. Provide shared workspace, a shared browser, etc. Enables users to communicate via chat and discussion forum services. Also includes mirrors, caching and archiving services, and virtual space management and security.

The registry also depends on a knowledge structure called the **ontology**, which defines relationships among words and concepts and supports rule clarification and glossary and other help services. Additionally, a set of **rules and constraints** that set parameters and adjudicate ambiguities are envisioned. These include consortium governance rules (e.g., how universal access laws are embodied; whether and how contributions are attributed; what restrictions apply to added value services offered to users who log on publicly; and what kinds of 'branding' information accompanies repository and primary objects) and standards such as units of measure, conversion formulas, and how adjustments are handled. The rules may also be specific to tools (e.g., how updates are handled and audit trails for error corrections; what kinds of scaled data operations are allowed).

Using this architecture as a context, we have made progress on several fronts toward realizing this vision. Note that in the figure, the objects and actions in bold represent work that we have addressed thus far and the grayed out objects and actions will be addressed in the coming year. Several interface threads advanced this year: The Relation Browser interface was revised to support string search and a user study demonstrated its efficacy for simple lookups as well as more complex data mining tasks. We are currently working with FedStats to mount the Relation Browser in a test area. We have also advanced the PairTrees designs that provide more power to people to visualize partitions of datasets on multiple feature sets in parallel. Work also continued on spatial audio for maps. We made substantial progress on the aiding understanding through help functions. The animated glossary work has developed a template for creating and maintaining animations. We are working with NCHS to install some example animated glossary entries on the NCHS site. We have also advanced our guidelines for multilayered help through animated demonstrations for TreeMaps.

In the Fall of 2003 we hosted a two-day metadata workshop that informed our thinking about statistical metadata. Based on this workshop, we developed a project DTD that is based on the DDI standards. We are working with BLS to formalize and evaluate the DTD by marking up several kinds of statistical objects. In addition to the metadata work, we have continued to develop a statistical ontology that will inform the metadata and animated glossary.

Another important thread of work that has advanced this year is our effort to crawl web sites and automatically categorize pages into facets that can be represented in the Relation Browser and other interfaces. We have investigated a number of machine learning techniques and settled on a hybrid approach that is instantiated in a toolkit that we hope agencies will be able to implement themselves in the years ahead. See <http://www.ils.unc.edu/govstat> for details on this work.