Designing a Metadata-Driven Visual Information Browser for Federal Statistics

Bill Kules and Ben Shneiderman

Department of Computer Science, Human-Computer Interaction Laboratory,
and Institute for Advanced Computer Studies
University of Maryland at College Park
College Park, MD 20742
{wmk,ben}@cs.umd.edu

Abstract

When looking for federal statistics, finding the right table, chart or report can be a daunting task for anyone not thoroughly familiar with the federal statistical system. Search tools help, but differing terminologies within the statistical agencies and a lack of familiarity of terms by information seekers limit their effectiveness. The FedStats Browser is a design for visually browsing federal agency statistical products and publications, using techniques that allow users to reformulate queries and iteratively refine results via simple, reversible actions with immediate feedback. This paper also discusses the characteristics of metadata needed for such a browser and the challenges inherent in acquiring that metadata.

1. Introduction

More than 70 federal agencies produce statistical tables, charts, reports and datasets and much of this information is placed on the World Wide Web for the use of researchers, practitioners, policy analysts, academics, and the general public. Simply finding the right table, chart, report or concept can be a daunting task for anyone not familiar with the federal statistical system. The Fedstats web portal helps information seekers take the first step in their search for federal statistics by providing an index of links to 700+ web sites. But then users are on their own, forced to navigate the disparate information structures of each web site to find documents of interest. Because web sites have historically been structured by agency or program, relevant documents may be scattered across several sites. Search tools such as Google can help, but differing terminologies within the statistical agencies and a lack of familiarity of terms by information seekers limit their effectiveness.

Several initiatives are attempting to address this problem by developing common metadata models and machine interfaces to support a unified user interface for information seekers. This work is often linked with efforts to implement enterprise content management systems. When fully populated, these systems could contain tens or hundreds of thousands of entries. Search interfaces similar to the Library of Congress online catalog[1] or PubMed[2] will be useful for users that have well defined queries and understand the statistical domain, but others with less familiarity of the domain will need more guidance and more flexible ways to find relevant documents and achieve their objectives.

This paper presents a design for visually browsing such large information spaces. It provides controls such as checkboxes, sliders and small image maps that enable users to progressively narrow their queries via simple, reversible actions that produce immediate feedback (under 100 msec). It takes advantage of the fact that different sub-domains have different metadata models with varying attributes by dynamically presenting controls that are applicable at different stages of an information-seeking task.

---

[1] See http://catalog.loc.gov.

[2] See http://www.ncbi.nlm.nih.gov/PubMed.

2. Related Work

Before describing the FedStats Browser, we briefly survey related work in the fields of visual information browsing and statistical metadata. Visual information browsing often starts with common one and two-dimensional layouts such as the FedStats index (which is organized like a typical book index) and the BLS home page[3] (which organizes 121 links into 19 thematic topics). (Allen, 1995) describes two systems for navigating and searching large collections of document records based on the Dewey Decimal System and the ACM Computing Review. FilmFinder (Ahlberg, 1993) uses a starfield display and interactive controls to enable users to rapidly browse and filter a dataset of films. The dynamic query techniques it introduced support progressive refinement of queries and immediate reversibility of actions. The Library of Congress Collection Browser (Marchionini, Plaisant et al., 1998) applies dynamic queries to provide a visual overview of library holdings. Users can adjust widgets for timeline and categories to filter the displayed list of collections. Flamenco (Hearst, Elliot et al., 2002) provides access to multi-faceted document collections using simultaneous menu techniques, as well as query previews. It builds on work on simultaneous menus reported in (Hochheiser and Shneiderman, 1999) and was influenced by the Epicurious web site[4], as reported in (English, Hearst et al., 2001). Dynamic queries are used in (Eaton, 2001) to organize web search results.

There have been several efforts to support integrated views of statistical data within the state and federal statistical community, academia and other countries' social science agencies. Most have focused on microdata, with some attention to aggregate data and little attention to higher-level products such as charts or reports (Gillman and Appel, 1997) (Mechanda, Johanis et al., 2003) (Leighton, 2002).

DataFerret[5] provides access to Census, BLS and other agency microdata and aggregate data. Using this tool, users can explore microdata datasets and extract variables or create their own tables of aggregate data. It provides some search capabilities for users to find concepts and variables**.** It does not provide variable descriptions or other metadata for users, nor does it index higher-level documents such as charts or reports. Nesstar Explorer[6] allows users to search across multiple catalogs of data products and within specific fields such as title, abstract, keywords, and date ranges. After finding a dataset, users can review the codebook, perform analyses, or extract individual variables. It works with data archives that are compliant with the Data Document Initiative (DDI) metadata standard (Ryssevik, 2001). Both of these tools require that the user have a well-defined query to specify the initial search criteria.

The Relation Browser (Marchionini, 1999), a research tool, applies dynamic query techniques and query previews to a set of 194 federal statistic web sites. Its interface provides an overview of topics and enables users to explore the set of web sites, filtering on a small set of attributes that includes topic, date, data type, and region. Users can filter web sites using these attributes, preview a list of sites, and then click on the desired site to open it. Visual feedback indicates the number of sites that satisfy attributes.

3. The Browsing Strategy

Browsing differs from search by allowing users to immediately experience the information space without formulating an initial query. Non-experts lack domain knowledge that is required for searching successfully (Marchionini, Plaisant et al., 1998). They do not know specific terms used to catalog documents; nor do they know the structure of taxonomies that partition the information space. This makes

---

[3] See http://www.bls.gov.

[4] See http://www.epicurious.com.

[5] See http://www.thedataweb.org.

[6] See http://www.nesstar.com.

it difficult for them to construct good queries, and their searches often return results with zero hits or far too many (Tanin, Plaisant et al., 2000). Browsing allows these users to gain an overview and learn how the information space is structured. Even for domain experts, if the problem is not clearly defined, browsing can help users clarify the problem and develop a strategy for solving it.

Visual information browsing emphasizes rapid filtering to reduce result sets, progressive refinement of search parameters, continuous reformulation of goals, and visual scanning to identify results (Ahlberg, 1993). Direct manipulation, dynamic query, and query preview techniques are used to provide immediate feedback to users as they explore. Actions are immediately reversible to encourage exploration.

4. Design of the FedStats Browser

The FedStats Browser design extends the original Relation Browser by providing a richer visualization and control set. It exposes more attributes to the user and allows the user to manipulate the attributes to successively narrow the search space. The two browsers are complementary – ultimately they could be unified in a layered interface that allows users to choose the power level of complexity appropriate for their task. As users become more comfortable with the interface they could select more powerful features.

Since the interface objectives include providing an overview of the collection, the left side of the interface is devoted to a hierarchical category list, initially showing all top-level categories (see figure 1). Users can drill down to sub-categories by clicking on the plus (+) symbol. Selecting a category immediately updates the document list and displays controls for additional filtering. Figure 2 shows the document list narrowed to 26 items after selecting Deaths from the category hierarchy and checking the By Gender and By Race demographics boxes.

The current set of filtered documents is displayed in the upper-right corner as a simple list. The current size of the list is always displayed above it, which allows the user to easily determine when the list is small enough to linearly scan. The list contains one document title per line. If the user places the pointer (hovers) over an item for 750 msec, a tooltip is displayed containing the categories under which the document is indexed, the time period covered, the spatial unit (e.g. states, region) and observational unit (e.g. individuals, institutions). A preview is displayed by clicking on the title. These two features support rapid relevance judgments by the user.

A set of common attributes is exposed in the dynamic query controls for filtering documents. For example, breakdowns by demographic attributes such as gender, race and income are often found in tables, so checkbox controls were created for each. Similarly, a double-ended slider is used to select a time period of interest, and a small image of the U.S. enables users to select by region. Additional controls are displayed when certain sub-categories are selected. For example, several behavior/risk factor attributes are relevant to the Health category, so checkboxes are displayed for them when Health or any of its subcategories are selected. To reduce the initial visual complexity for users, all attribute controls are initially hidden. When a topic in the hierarchy is first selected, they are then displayed.

For both the attribute controls and the category hierarchy, query previews in the form of parenthesized counts provide an indication of how many of the currently filtered documents fall in that category or have the attribute. Zero-hit queries are avoided by de-activating (graying out) checkboxes for attributes that would result in no matches. Tooltips are used on the controls to provide additional description for terms that may not be familiar to users.

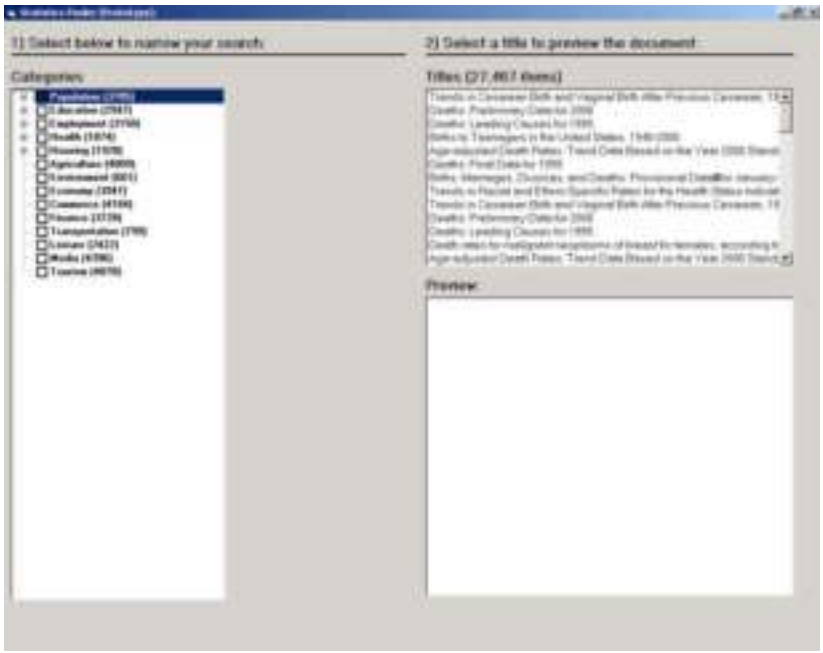

Figure 1. The initial screen, showing all top-level categories and all documents. Additional controls are hidden to minimize interface complexity.

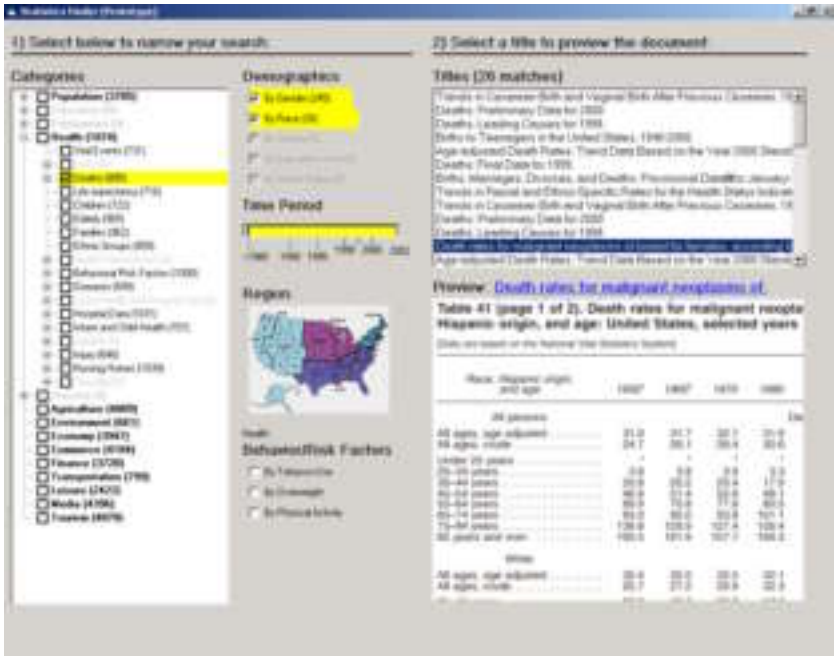

Figure 2. After user selects the Deaths category and checks two demographics boxes, the document list is reduce to 26 items. Clicking on a title, previews that document, a table of death rates containing the desired gender and race breakdowns.

5. Metadata Requirements

This browser is intended for large, heterogeneous document collections (on the order of $10^4 – 10^6$ documents) with rich metadata, i.e. with a large number of attributes for each document. Multiple classes of documents will need to be indexed. At the finest level of granularity will reside concepts and variables

from individual studies or aggregate tables from such studies. This will be appropriate for certain end-users (even non-experts) who want to answer specific questions or make specific comparisons. Many end-users who are interested in higher level, less well-defined questions, may prefer to find reports, charts, or summary tables. Indeed, even users with similar objectives will have different styles, and one user may want a narrative report while another wants to compare the numbers herself. The metadata model has to accommodate this level of diversity while also being integrated across agencies (Marchionini, Hert et al., 2000) and (Hert, in press).

The model must also have a flexible way to specify relationships between cataloged items. For example, a single survey could yield cataloging records for the concepts, the aggregate tables and a summary or report, and the records must capture these interrelationships. This will be needed for end-users, so they can drill-down for more details or move up for more general information. It will also be important for maintenance and administration of such a collection, so that – for example – a survey author can quickly review and update all cataloging records when updating a published document.

A review of National Center for Health Statistics (NCHS) and Center for Disease Control (CDC) documents produced the following potential metadata elements:

- Caption for Figure/Table
- Topic/Category
- Keywords
- Title of Containing Document
- Abstract for Containing Document
- Publication Date
- Dates covered by data (could be single year, range or discontinuous; some data goes down to months)
- Document Type - Report, Table, Figure, Summary, "Highlight", Press Release, web page, etc.
- Document Format - PDF, HTML, etc.
- Whether the item provides breakdowns by:
  - Gender
  - Race
  - Age
  - State
  - Geographic Region
  - Behavioral characteristics, e.g. smoking
  - Education Level
  - Income Level
  - Marital Status
  - Diagnostic Category

Existing metadata models do not support all of these attributes, and agencies face several challenges when attempting to incorporate such metadata. The existing standards do not model this metadata effectively, and so lack the relevant attribute fields that could be populated. Agency publication processes rarely include a rigorous cataloging step, so there is no current role to assign the tasks to. Few of the statistical agencies have as their mandate to produce detailed metadata or cataloging data, and it is difficult for them to allocate resources to the task, even though the advent of web publishing is pushing them to put ever increasing amounts of material on their web sites. A more subtle challenge is that survey designers and analysts often do not have responsibility for publication, and are not aware of the broad diversity of users.

6. Conclusion

The visual information browsing design described in this paper has the potential to significantly simplify the task of finding federal statistics for both experts and non-experts. When combined with complementary information retrieval techniques like search, book marking, and history keeping, it will enable users to quickly overview, zoom, filter, and evaluate potential documents without needing a pre-existing knowledge of the information architecture, terminology or keywords.

We are now building a prototype based on this design, and will evaluate it in collaboration with our agency partners. With our agency partners, we are also tackling the significant challenge of developing and populating archives that incorporate the rich set of metadata needed for sophisticated filtering and browsing of statistical documents.

Acknowledgements

This material is based upon work supported in part by the National Science Foundation under Grant Number EIA 0129978 (see also http://ils.unc.edu/govstat/) and the National Center for Health Statistics.

References

Ahlberg, C., Shneiderman, B. (1993). Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays. *ACM CHI '94 Conference Proc. (Boston, MA)*: 313-317.

Allen, R. (1995). Two Digital Libarary Interfaces That Exploit Hierarchical Structure. *DAGS95: Electronic Publishing and the Information Superhighway*.

Eaton, C., and Zhao, H. (2001). Visualizing Web Search Results. *Available at http://www.cs.umd.edu/class/spring2002/cmsc838f/Project/QueryResults.pdf*(11/26/2002).

English, J., Hearst, M., Sinha, R., Swearington, K. and Yee, P. (2001). Examining the Usability of Web Site Search. *Available at http://bailando.sims.berkeley.edu/papers/epicurious-study.pdf*.

Gillman, D. and Appel, M. (1997). The Statistical Metadata Repository: an electronic catalog of survey descriptions at the U.S. Census Bureau. *IASSIST Quarterly* **21**(2): 34-51.

Hearst, M., Elliot, A., English, J., Sinha, R., Swearingen, K. and Yee, P. (2002). Finding the flow in web site search. *Communications of the ACM* **45**(9): 42-49.

Hert, C. (in press). Supporting End-users of Statistical Information: The Role of Statistical Metadata Integration in the Statistical Knowledge Network. *Proceedings of the 2003 National Conference on Digital Government Research*.

Hochheiser, H. and Shneiderman, B. (1999). Performance Benefits of Simultaneous Over Sequential Menus as Task Complexity Increases. *International Journal of Human Computer Interaction* **12**(2): 173-192.

Leighton, V. (2002). Developing a new Data Archive in a Time of Maturing Standards. *IASSIST Quarterly* **26**(1): 5-7.

Marchionini, G. (1999). An Alternative Site Map Tool for the Fedstats Statistical Website, School of Information and Library Science.

Marchionini, G., Hert, C., Liddy, E. and Shneiderman, B. (2000). Extending understanding of federal statistics in tables. *Proceedings of the 2000 Conference on Universal Usability*, ACM Press. 132--138.

Marchionini, G., Plaisant, C. and Komlodi, A. (1998). Interfaces and Tools for the Library of Congress National Digital Library Program. *Information Processing & Management* **34**(5): 535-555.

Mechanda, K., Johanis, P. and Webber, M. (2003). Conceptual Model for the Definitional Metadata of a Statistical Agency. *Open Forum 2003 on Metadata Registries*, Santa Fe, NM, USA

Ryssevik, J. (2001). The Data Documentation Initiative (DDI) metadata specification. *Available at http://www.icpsr.umich.edu/DDI/PAPERS/.*

Tanin, E., Plaisant, C. and Shneiderman, B. (2000). Browsing Large Online Data with Query Previews. *Proceedings of the Symposium on New Paradigms in Information Visualization and Manipulation (NPIVM) 2000*, Washington D.C, ACM Press