E-Tables: Non-Specialist Use and Understanding of Statistical Data

Gary Marchionini, University of North Carolina at Chapel Hill
Carol Hert, Syracuse University
Ben Shneiderman, University of Maryland at College Park, and
Liz Liddy, Syracuse University
[march@ils.unc.edu; cahert@syr.edu; ben@cs.umd.edu; liddy@syr.edu]

Abstract

This paper provides a progress report on a project that aims to understand how people think about and understand statistical data presented in tables. The focus is on people who are not statistical specialists. To this end, studies of various user populations were conducted, underlying data enrichment and explanation was investigated, and interactive user interface prototypes were developed and tested.

1. Introduction

Statistical data is an important type of information that is finding growing importance among non-specialists who access this data via the WWW. Although great advances have been made in electronic tools for collecting data (e.g., computer assisted survey information collection methods [Dippo, 1993]), managing and manipulating small data sets (e.g., spreadsheet packages), and managing and analyzing large data sets (e.g., database management and statistical packages, respectively), dissemination of statistical data to general populations has generally been treated as a text transmission and display problem. Electronic text has been studied by researchers ranging from the earliest text processing systems through hypertext rhetoric. Likewise, electronic text has been adopted in practice by people everywhere through word processing and email. We believe that electronic statistics represent a distinct medium and require research and development attention. Ultimately, statistical data sets are like spatial data sets—structured relationships among well-defined attribute sets (e.g., variables) that may be partitioned, aggregated, and rendered in a variety of ways to suit the needs of users. We have focused on one important type of rendering, electronic tables (e-tables). Beginning with exemplars of tables that are produced by government agencies, we have investigated how people think about and use these tables, developed prototype user interfaces to help people work with e-table equivalents, and studied how the underlying data may be structured in order to aid user understanding and use. Over the past 18 months, our research questions have evolved from fairly specific questions about interface design to more general questions about e-tables. These questions include: What is an e-table? How are e-tables different that other electronic media? What do people want and need to do with e-tables? What kinds of problems emerge as non-specialists begin to use e-tables? What kind of system features will assist people in making better use of e-tables?

This paper presents preliminary results of a collaborative effort to understand e-tables and their use by non-specialists, design prototype user interfaces to promote their use, and explore issues of table production and distribution in government arenas. There are three main clusters of results related to users, data, and systems.

2. E-table Users.

Hert and Marchionini (1997; 1998), used a variety of methods, including interviews of users and agency personnel, transaction log analysis, content analysis of email messages, and usability studies, to develop an understanding of user tasks, user expectations, how intermediaries help users, how users search statistical websites, and the organizational aspects of providing information via the web. Hert (1998) observed user-intermediary interactions and asked intermediaries about the questions they asked users as they worked to understand a user's inquiry. Building upon this work, the current project team conducted focus groups with data users to ascertain their concerns and problems with tabular representations; conducted workshops with senior citizens to understand their specific barriers to using tables; observed users searching for statistical information and via think-aloud protocols identified uncertainties, problems, questions that they had; performed eye-tracking studies of users of both paper and electronic versions of tables; and conducted usability tests of the interface prototypes. These efforts yield a rich sketch of how non-specialists think about and use statistical tables. We have learned that in general, non-specialists lack knowledge of survey methodology, lack an understanding of the structure of a domain and the people and information entities (both metadata and key publications that disseminate the information) within it, and lack necessary information handling and technical literacy skills. These deficiencies often lead to mismatches between information needs and available information and frustration as people look for statistical data. It is important to note, however, that these deficiencies are not complete—people often have sophisticated understandings of particular aspects of data and these sophistications complicate design specifications based upon minimal sets of anticipated user skill. For example, some of the senior citizens interviewed did not understand statistical concepts of error but well-understood that morbidity data for certain diseases cannot be fully accurate given that it comes from death certificates and the many reasons that cause of death may be an estimation or simplification of reality (e.g., multiple causes, insurance constraints, etc.). In another interview, a subject questioned a state by grade table of gasoline prices that distinguished self-serve from full-service costs, noting that in New jersey, self-serve is illegal. These various data sources show that people may have sophisticated understandings of very specific and personally relevant data while not having the statistical language or generalized concepts that provide entry to those details from the masses of data available.

Our studies of users also focused on interface issues. Interviewees noted difficulties with scrolling, comparing data across columns or rows, understanding row or column headings, and size of fonts. The eye tracking study (Xu, 2000) showed that large amounts of eye movement (up to 40% of all eye movement) is devoted to vertical moves between cells and column headings. Most importantly, user data reinforces the need for several layers of on-demand explanation and elaboration for tabled data.

In addition to specific interface issues, the user studies also aimed to identify particular questions that people ask in order to understand specific table data. A preliminary set of questions have been cataloged and answers sought in order to inform overall data and metadata flow. In some cases, answers are possible with simple definitions, in others, footnotes provide the answer, in other cases, the answer is available in texts or reports in which the table is presented, and in still other cases, the answer required personal explanation by personnel at the agencies producing the data. For the different questions and answer categories, specific metadata requirements and interface recommendations are specified. See Hert et al (in press) for preliminary examples of these mappings from question to answer to system designs.

3. Statistical Data

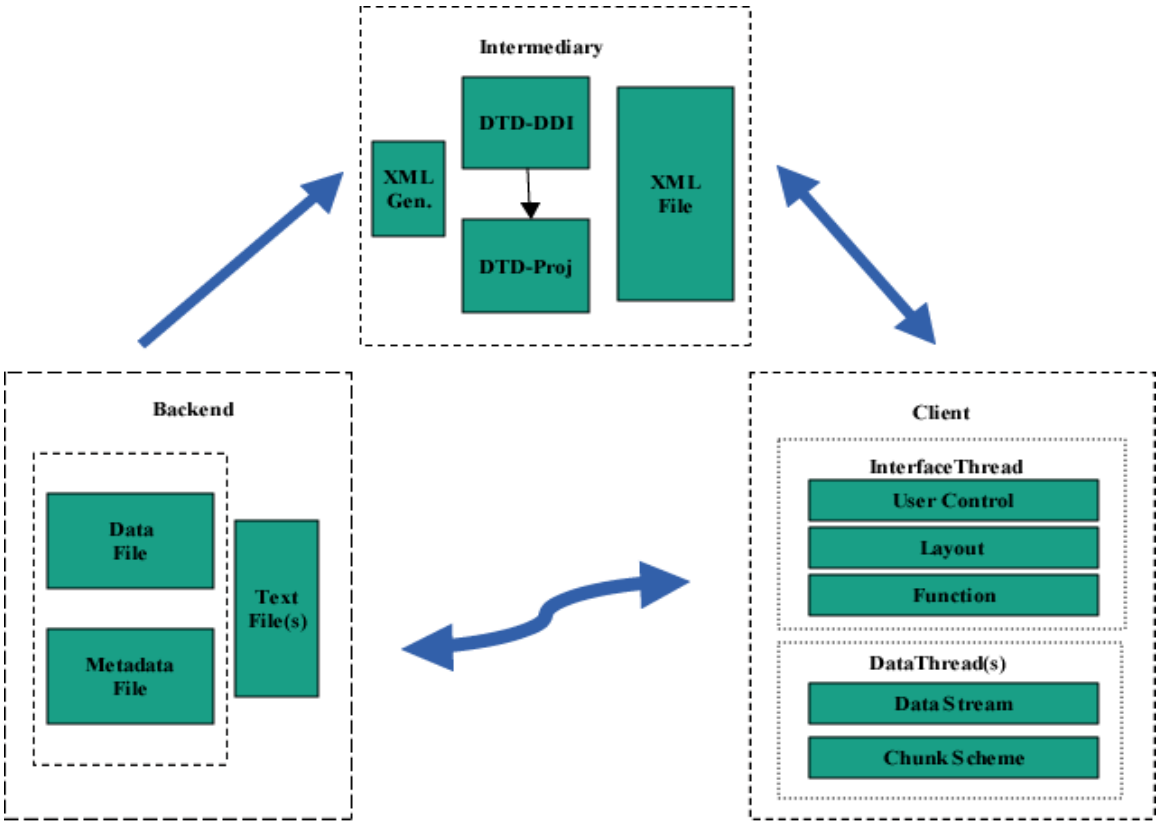
Statistical data is collected for a variety of purposes by the federal government and through a wide range of channels (e.g., in-person surveys, phone surveys, required reports on a time or event driven basis, etc.). In the U.S., there is no central statistical agency and more than 70 different agencies have dedicated statistics gathering mandates. Different agencies (and different departments within large agencies) use different techniques and systems for collecting and managing these data. A variety of survey instruments and coding schemes are used to collect and process data. Different document management and database management systems are used for microdata files that are in turn used by staff or confidentiality-cleared researchers. These staff and researchers may use different statistical packages to process the data and produce macrodata files and various textual reports. These files and reports are then made available in a variety of formats and systems (e.g., print publications, HTML files, PDF files, ASCII text files, etc.). Along the way, any specific variable may acquire or lose associated "metadata." Codebooks for microdata or data dictionaries for database tables typically do not migrate from one system to another as data flows from collection to dissemination. Although some explanations of data are provided in the text accompanying tables, in WWW presentations, these texts may be discretely separated (e.g., in different windows) or unavailable.

As part of our goal to create e-tables that non-specialists can understand and use, we recognized the need to incorporate various types of metadata, explanations, and access points directly on the e-table. Although the electronic medium provides new opportunities for layering such data and making these data interactive, understanding what metadata is available, what is useful to users, and how it can be made available to users in an on-demand fashion are issues that require investigation. To this end, the team worked to identify a metadata standard, create XML documents that contained and identified these metadata, and map these XML documents to interfaces. The interviews generated questions that were categorized into classes that in turn were associated with metadata types. For example, many questions required variable definitions or elaborations and so we included definitions/elaborations as metadata elements and in turn mapped these to interface mechanisms such as mouse over pop-up (tooltips). The Data Documentation Initiative (DDI) model for metadata was adopted (http://www.icpsr.umich.edu/DDI) to guide a project specific document type definition (DTD) for XML markup. A Java script was developed to automate XML markup from database files for a few very specific tags (e.g., row or column headings, table title, etc.), but for the prototype, several example tables were marked up by hand. A hierarchical model was adopted that embedded cell level entries within column/row entries within document structure (document description, study description and data files description). These metadata were then mapped onto interface mechanisms. For example, in a gasoline table, at the cell level, a mouseover action caused a tooltip to pop up that elaborated what that cell (1.707) means-- $1.707, estimated price per Gallon (including taxes) in PADD 1A Conventional Areas on Dec.25, 2001. A single click on the cell yielded a set of metadata (in a pane reserved for this purpose) for term definition, study information (source), and contact information (an email address). Likewise, unique information that applied to the entire table (e.g., information that might go in table title footnotes or commentary in accompanying text) for a table was mapped to the table title.

Figure 1 provides the overall architecture for the system. In the figure, the backend files are linked to the client through intermediary services that depend on XML markup as the bridge to

the client-side interface. In this project, a java applet parses the XMl and renders the interface for use. The multithreaded client architecture is detailed in Mu and Marchionini (in press).

Figure 1. System Architecture.



4. User Interfaces

We wanted to give people multiple ways to access and understand tables. Figure 2 shows these options. On one hand, the information seeker may enter a natural language query that returns a set of tables that is selected and displayed by the user in a table browser. Another option is for the information seeker to explore the data space using dynamic query interface overviews that allow them to dynamically define a table of interest and then display and manipulate the table in the table browser. Alternatively, the users may use a hierarchical selection approach (like a file directory system) to select tables that are then displayed and manipulated in the table browser.

The query module is described in Liddy & Liddy (in press) and the visual overview module is described in Tanin et al (2000), Hert et al. (in press) and Marchionini et al., (2000). The table browser gives users basic spreadsheet-like features such as selecting and manipulating cells, rows, or columns; changing column orderings (drag and drop); simple computations (e.g., mean, max, min, percentage); and freezing column and row headings so they do not scroll off the screen. Password protection, several subtable creation options, and simple zooming are also provided. Most importantly, the browser provides cascading levels of metadata through mouseover and single click control mechanisms (see Mu & Marchionini, in review; Marchionini et al. 2000 for details and screen displays.

Figure 2. Information Seeker Options

5. Conclusion

This abstract provides a high level overview of our investigation of e-tables and prototype designs for interfaces that will help non-specialists find, understand, and use statistical information. See http://istweb.syr.edu/~tables/ for papers and background on the project. See the table preview demos at http://www.cs.umd.edu/~egemen/demo.html for examples of visual overview prototypes. See http://idl66.ils.unc.edu/table/current for table browser demo.

Acknowledgements

This work is supported by National Science Foundation Grant Number 9876640. Fred Gey provided metadata assistance and several graduate students work on this project: Simon Mu, Zhen Zhen Deng, Egemen Tanin, Naybell Hernandez, and Kristin McKeown-Armstrong.

References

Dippo, C. (1993). Designing an Instrument For Computer-Assisted Data Collection in the Current Population Survey. Bureau of Labor Statistics Research Reports. http://www.bls.gov/orersrch/cp/cp930010.htm

Hert, C.A. (1998); "FedStats Users and Their Tasks: Providing Support and Learning Tools: Final Report to the United States Bureau of Labor Statistics", Available at: http://istweb.syr.edu/~hert/BLSphase2.html

Hert, C.A. and Marchionini, G. (1997); "Seeking Statistical Information in Federal Websites: Users, Tasks, Strategies, and Design Recommendations: Final Report to the Bureau of Labor Statistics", http://ils.unc.edu/~march/blsreport/mainbls.html

Hert, C., Marchionini, G., Liddy, L., & Shneiderman, B. (in press). Extending Understanding of Federal Statistics in Tables: Integrating Technology and User Behavior in Support of Universal Usability. Interacting whab1.3(1.142c.ed)1lq3 11j-T11.9(i)8

Hert, C. & Marchionini, G. (1998). Information seeking behavior on statistical websites: theoretical and design implications. Proceedings of the 61st Annual Meeting of the American Society for Information Science. (Pittsburgh, PA, Oct. 25-29, 1998). P 303-314.

Liddy, E.D. & Liddy, J.H. (In Press); "An NLP Approach for Improving Access to Statistical Information for the Masses"; Proceedings of the FCSM 2001 Research Conference.

Marchionini, G., Hert, C., Liddy, L., & Shneiderman, B. (2000) Extending Understanding of Federal Statistics in Tables. ACM Conference on Universal Usability. (Washington), 132-138.

Mu, X. & Marchionini, G. (in press). An Architecture and Prototype Interface for an Online Statistical Table Browser. American Society for Information Science and Technology Annual Conference (Fall 2001).

Tanin, E., Beigel, R., and Shneiderman, B.; (1996); "Incremental data structures and algorithms for dynamic query interfaces"; ACM SIGMOD Record 25, 4 (December 1996), 21-24, and Proc. Workshop on Information Visualization, (November 1996).

Tanin, E., Plaisant, C., and Shneiderman, B., Browsing Large Online Data with Query Previews, Proceedings of the Symposium on New Paradigms in Information Visualization and Manipulation - CIKM, 2000, USA

Xu, A. Eye Tracking Study on Identifying and Analyzing User Behavior - Eye Movements, Eye Fixation Duration and Patterns - When Processing Numeric Table Data in Paper or PDF Format. Master's Paper, School of Information and Library Science, UNC-Chapel Hill. Available at http://ils.unc.edu/idl/details/AirongXu.pdf