Extending User Understanding of Federal Statistics in Tables

Gary Marchionini[1], Carol Hert[2], Liz Liddy[2], and Ben Shneiderman[3]

Interaction Design Laboratory[1]
University of North Carolina at Chapel Hill
march@ils.unc.edu

School of Information Studies[2]
Syracuse University
{cahert,liddy}@mailbox.syr.edu

Human Computer Interaction Laboratory & Dept of Computer Science [3]
University of Maryland
ben@cs.umd.edu

ABSTRACT

This paper describes progress toward improving user interfaces for US Federal government statistics that are presented in tables. Based on studies of user behaviors and needs related to statistical tables, we describe interfaces to assist diverse users with a range of statistical literacy to explore, find, understand, and use US Federal government statistics.

Keywords

Statistics, dynamic queries, tablular data, data exploration, user interfaces

INTRODUCTION

As the World-Wide Web (WWW) reaches larger portions of the world's population interface designers must consider the needs of more diverse users. Gone are the days when motivated users adapt themselves to whatever interface is provided in order to achieve their goals. Nowhere are the needs to create interfaces that meet the needs of the maximum portion of the population greater than in government websites. Governments at all levels exist to meet the needs of citizens, and as services and resources move to the WWW more citizens are required to use information technology to take advantages of those services. At the US Federal level, citizens and others obtain (and file) tax forms through Internal Revenue Service websites, social security recipients obtain and provide information through websites, public interest groups follow legislation through the Library of Congress Thomas website, and people find health information through the US National Institute of Health's websites. One area of government information that takes on increasing importance in an informated society is government statistics that are used to inform decision-making in a variety of personal and corporate contexts.

Because the statistical literacy of the citizenry tends to be low (Moore, 1997), the problems of digital divide and equitable access are exacerbated for government statistics. A key goal of interface design is to help users understand the data behind highly distilled statistics, usually presented in tabular forms. This paper presents a progress report of work that aims to create improved interfaces for government statistics found in the many types of tables produced by more than 70 US Federal government agencies.

Our main goals are to broaden and improve citizen's and others's understanding and use of statistical tables. One approach to broadening access is to provide multiple interfaces that serve a wide range of user expertise and working styles. Improving understanding requires interfaces that provide rich explanations for statistical concepts and easy-to-use tools for finding, viewing, manipulating, and extracting statistics. Our general approach to broadening is to define sets of alternative yet coordinated views from which citizens can choose. Our approach to explanations is to provide on-demand, cascading sets of metadata and elaborations. Our approach to tools is to extend interface techniques found useful in dedicated statistical environments such as spreadsheets, statistical packages and WWW-based environments and to augment these techniques with additional exploration and access techniques. Each of these approaches aims to empower individuals to define and control their interactions—i.e., we take a strongly user-centered design philosophy to the overall project.

To this end, our first subgoals are to understand what people already do with statistical data and investigate how statistics may be used by broader communities of citizens. We conducted a series of interviews and focus group

sessions to understand citizen needs and behaviors. A second set of subgoals relate to designing and testing interfaces that link the vast data (surveys, reports, datasets) in federal systems to the needs and capabilities of citizens and others. To achieve these subgoals, we are developing an interface grammar that relates basic statistical objects to user actions. Simultaneously we are creating prototype interfaces that help people explore statistical collections, find specific tables, understand their contents, and use them to meet their needs.

PROJECT CONTEXT

Our current work (funded by the National Science Foundation's Digital Government Initiative) is focused on statistical tables. There are several rationales for this focus. Although there is a substantial effort given to graphical representations of data (e.g., Carr, 1998; Wainer, 1997; Wilkinson, 1999) tabular display treatments are treated minimally at best (e.g., massive volumes such as Brinton, 1914 address graphics methods while early volumes on tables tend to be well under one hundred pages, e.g., Hall, 1943; Walker & Durost, 1936). Tables are a common conceptual and presentational structure by which statistical data are stored and represented (e.g., Tufte, 1983, p 178 has noted that tables are "clearly the best way to show exact numerical values."). Data in tabular form are used in analyses to generate graphic representations and analytic reports, as well as to contextualize specific values. However, they are often difficult to find, interpret and use. Most search engines do not retrieve data directly from statistical tables they often do not even identify tables within text). Once found, users must struggle to understand the highly distilled numbers and the meaning of column and row headers. Additionally, many of the ways in which these data might be used (e.g., comparisons, grouping) are difficult to accomplish in these static tables. The ubiquity of tables along with the associated challenges suggest that research into improvement of table retrieval, interpretation, and use has the potential to significantly improve access to data produced by statistical agencies around the world.

A huge range of tables is available. Different table forms may exist for data storage, data analysis and data presentation. Tables vary in their size. They also may represent the "raw data" with only minor summarization; full summaries or composites build from several or many raw data sets. This project is using several exemplars from a variety of US Federal agencies as the first step to providing guidance in how to improve access to tables. In particular, we are examining the tables presented in government statistical portals such as Fedstats (www.fedstats.gov) and the American Fact Finder (http://www.census.gov), and CD-ROM sources such as the Statistical Abstracts of the United States. In addition we are working with our agency partners under the Digital Government Initiative, including the Bureau of Labor Statistics, National Center for Health Statistics, Bureau of the Census, and Energy Information Agency.

A further impetus for this project is that with the accessibility of information from US Federal agencies via the WWW, an increasingly broad range of users can now access these data. While tables were once the domain of experts, agencies can now expect that their data will be found by high school students, senior citizens, and the public at large. These users are not likely to have the high level of statistical literacy, numerical literacy, and knowledge of the agencies that characterizes experts, thus multiple tools to facilitate access are critical (Hyland & Gould, 1998).

Although there have been several innovative attacks on specific aspects of the overall interface challenge, there is no integrated approach to the multiple problems of finding, understanding, and using tabular data. The Table Lens (Rao & Card, 1994) provides a rich set of interface mechanisms for focusing in on different regions of a table while maintaining the surrounding context, but does not address underlying metadata issues. The DEVise system (Livny et. al, 1997) allows users to visualize multiple tables in interesting ways if common attributes are available across all the tables. The TINTIN retrieval mechanism (Pyreddy & Croft, 1997) automatically detects tables in documents and supports user queries applied to the tables but does not provide an interface for end user support or tabular display. Statistical packages such as SAS and SPSS provide good display capabilities but do not link data displays to data dictionaries and assume considerable expertise on the part of users. None of these approaches provide user support for statistical literacy as part of search and exploration

CITIZEN USE OF STATISTICS AND TABLES

The work reported here is building on several years of work investigating information seeking and use of statistics. Hert & Marchionini (1998) used a variety of data collection techniques to gather information about the kinds of questions people ask about statistics at the Bureau of Labor Statistics and other federal websites, and how people use such statistics. Based on focus group interviews with intermediaries and users, content analysis of email requests to BLS staff, interviews with BLS analysts, and analysis of BLS website transaction logs, a taxonomy of tasks and user types was developed (see http://ils.unc.edu/~march/blsreport/mainbls.html for details).

User tasks were organized into three categories: pragmatic, semantic, and syntactic. The *pragmatic* category includes three subcategories: goal (learn something new, verify, judge/evaluate/compare, explore, referral/intermediate, ongoing, and planning/forecasting); task constraints (time, amount, geography); and system (appropriateness of database to task, location and extraction facilities, formats available, user preferred entry point into system or optimal path to information). The *semantic* category has four subcategories: topic, abstraction level (concrete/abstract), specification (specific/unspecific), complexity (faceted/non-faceted, number of facets). The *syntactic* category has three subcategories: expression type (what, where, who, how, and why); goal type (closed—known item, open/interpretive, and accretional); and specificity of expression (inclusion of information about the user's situation and about how that user interacts with the system).

User types include: business users (people using the sites in support of for-profit business activities), academic users (researchers, graduate students, faculty), the media (journalists), the general public (people using the sites in support of non-work related tasks), government users (from all branches of government, both international and national, down to local levels), education users (teachers and students from K-12 institutions), statisticians (both within government agencies and external to them), and users at libraries/museums, and other non-profits users (e.g., think tanks, special interest groups).

We are conducting scenario-based focus group interviews with sophisticated users (e.g., researchers) as well as citizen groups (e.g., senior citizens in retirement communities) to focus more specifically on statistical tables. Preliminary findings are that sophisticated users want more access to raw data—either microdata directly from surveys or more detailed aggregated data. If such access is not possible, they want tables that can be manipulated. Comparison is a common activity for all users (e.g., comparing income across counties or states). They also want and need context for tables, including term definitions, data quality indicators such as variances within cell values, and other metadata that supports data understanding and provenance.

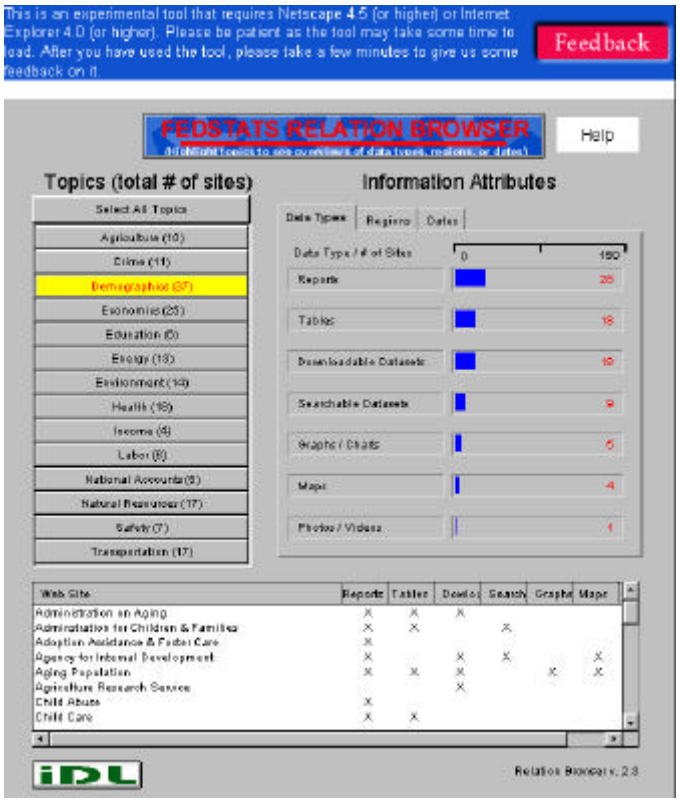
We continue to collect information on citizen needs for statistical data and feed these results into the interface prototypes. Our next activities will investigate what information (including metadata) are necessary to support exploration, interpretation, and use tasks.

EXPLORATION WITH OVERVIEW TOOLS

One approach is illustrated in the Relation Browser that has been in experimental use on the Fedstats website (a portal to all government statistics) since the fall of 1999. In this interface, users see an overview of what data types (reports, tables, downloadable datasets, searchable datasets, graphs/charts, maps, and photos/videos) and quantities (number of web sites) of data are available for different topical areas (fourteen topics defined by the Fedstats task force representing 70 agencies and 200 different statistical websites). Users may also choose to explore geographic and temporal attributes for these topics. As users see what is available for given topics (overview of those topics) previews of specific web sites within this partition are provided in the lower window, showing the website title and what data characteristics it provides (specific URLs are also available). In Figure 1, as the user passes the mouse over the demographics topics they can see that there are 37 websites that have demographic data , that 26 of those sites have reports and four have maps. The user can also see that a specific website titled Child Care contains reports and tables but no other types of data. Thus, users are able to gain an overview of the entire range of government statistics as well as previews of any of the 200 websites without clicking the mouse or waiting for sites to load. Reports of preliminary usability tests for two iterations of this tool are reported in Marchionini, et al. (2000).

We have demonstrated the efficacy of overviews of information spaces and previews for information objects in those spaces in a variety of other contexts (Doan et al, 1997; Greene et al, 2000). A key requirement for overview and preview interfaces is careful specification of metadata for the underlying data. We are exploring ways to gather and update such data automatically (site robots) and systematically (templates for webmasters) as well as ways to transfer this metadata efficiently since the metadata must be present on the client side to insure the dynamic query interactions.

Figure 1. Relation Browser Interface



EXPLORATION BY SEARCH PATH

In addition to using novel interfaces to gain overviews of statistical data that is available, people should be able to search for specific statistical tables. Tables present special challenges since search engines do not parse the internal elements of tables. We are developing a search query grammar that supports query interfaces for tables. We are currently modeling this grammar with the Economy at a Glance tables at BLS. The grammar is based on samples of queries posed to different agencies and the task taxonomy described above. Basic elements accommodated within this grammar are: location (where), condition (what), population of interest, time (when), and number/percentage/rate (how much). Questions such as "What percentage of people living in Alabama did not have jobs in the summer of 1999?" are represented in the grammar and table elements such as title, row headings, column headings, notes, and cell values (numeric values as well as units and other metadata associated with the values) are searched to match the grammar template.

One interface solution we are exploring to make this grammar more explicit to users without requiring some SQL-like syntax is to provide a set of natural language templates with pick lists for the where/what/who/when/how much elements. A more general interface would allow users to type a natural language query and use natural language processing and relevance feedback techniques to parse and clarify the query (Liddy, 1998).

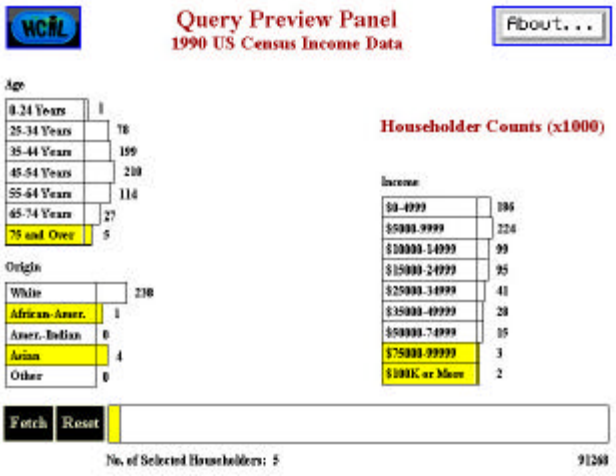
EXPLORATION BY QUERY PREVIEWS

Another approach to giving users an understanding of the contents of a statistical database is to use query previews that show the distributions of data (Shneiderman, 1998). We have built several prototypes for US Census data to allow citizens and others to explore data by clicking on key data attributes and see the size of resulting database partitions. Figure 2 illustrates such an interface where the user has selected a partition by selecting an age group (75 and over), two racial categories (African-American and Asian) and two income ranges (75K-100K and over 100K). The size of the result set is shown visually to be approximately 5,000 households. By showing the cardinality of the result set users gain an understanding of the data distribution and avoid posing zero-hit or mega-hit queries. Posing such a query in SQL would be quite challenging for most citizens but clicking and immediately seeing the number of households meeting these conditions is easy, immediate, and visually apparent (see Tanin et al. 2000, for details on this work).

INTERPRET AND USE THE DATA

Once users have located a table, whether through a drill-down exploration or a search query result that extracts the data, they should have easy control over viewing and using the table. People must be able to read, navigate, and manipulate tables and the data in tables. There are three sets of challenges to reading and understanding a table. The first set of challenges relates to user experience and needs. This includes issues such as understanding dataset concepts (what the records represent, what terminology means), understanding general statistical concepts (e.g., sampling, weights), managing the high cognitive load requirements inherent in a large set of numbers,

Figure 2. Interface for Previewing Population Statistics



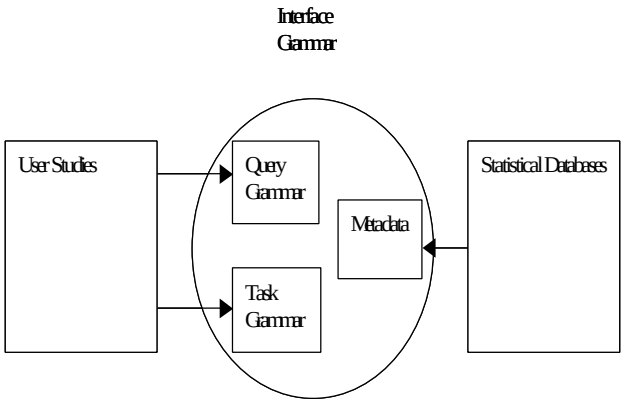
understanding specific values (what they mean, what can be done with them), and understanding where data come from (both for understanding and for citation).

The second set of challenges relates to the constraints on data itself and include documentation/metadata availability and consistency, data formats may not be what users need (e.g., comma delimiters, color settings), data may not be backward compatible with existing data, and data may not be sortable or manipulable. Finally, there are challenges related to the online medium such as on screen displays do not look like/act like paper, and screen real estate limits what can be seen.

The third set of challenges relates to navigation (e.g., scrolling loses headers and stubs, notes are not easily accessible, cell values are not easily linked to variable names) and manipulation (e.g., people may combine or summarize data inappropriately, adjustments or revisions to data may not be apparent across cells).

To address these challenges, we plan to provide on-demand metadata annotations to all statistical objects and offer flexible user control mechanisms for viewing and manipulating those objects. Our approach is to develop an interface grammar that maps statistical table objects onto user actions (Figure 3). Studies of user behavior are the basis for a query and a task grammar specific to tables. Similarly, studies of the statistical data available from various government agencies are the basis for extracting metadata elements. These grammars are in turn mapped onto the metadata to define the user interfaces.

Figure 3. Empirically Defined Interface Grammar for Statistical Tables



This grammar is analogous to the graphics grammar developed by Wilkinson (1999) to render different graphical representations for a data set. Our interface grammar actions are meant to link to the query grammar, both of which are rooted in our user needs and behavior investigations. Thus, these grammars act as a conceptual interface between the varied data produced by federal agencies and the varied tasks and experiences citizens bring to federal statistics websites.

Statistical table objects include cell (the atomic unit, may have multiple layers, may have computational elements that support customization, security, and "telling its story"), row (has a header, may have multiple layers), column (has a header, may have multiple layers), header (modifies a cell, row), subtable (subset of a dataset or fixed table), table, survey (the source(s) of data represented in the table), and agency (the producer/owner of the data). Actions that users can take include:

- view (scroll, pan, zoom, mouse-over, jump),
- change view (resize, reorder, graph/visualize),
- define table (selection from menu, drag and drop, partition, find),
- calculate,
- help (tell story, tutorial, demo, reference), and
- print/save.

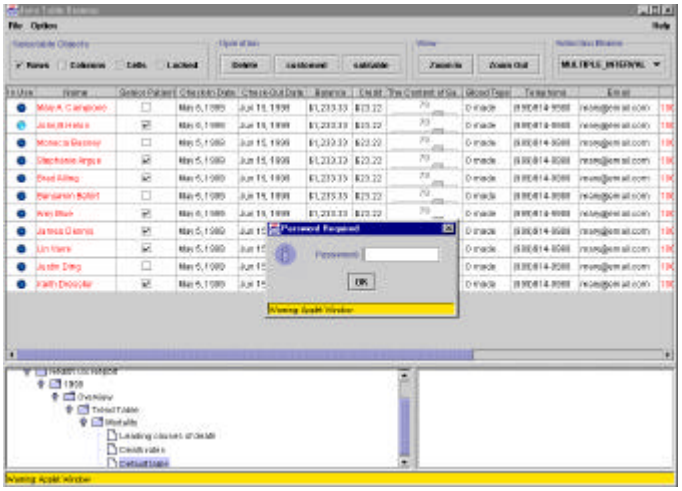
Our preliminary mockup is based on this interface grammar and supports many of the task actions while eliminating many of the limitations typically placed on web-based tables. (See Figure 4). The current prototype requires the Java 1.2 plugin and provides in the WWW environment some of the features available in spreadsheet and other local applications as well as some additional features that aid user understanding for federal statistics. These features include:

- The column headings do not scroll off the screen.
- The columns can be dragged around and exchanged and width adjusted.
- To keep the leftmost column frozen even when scrolling right, a lock button is provided
- The zoom in and out buttons allow closer or more distant looks (the present prototype provides a discrete zoom rather than a Jazz-like [Bederson & McAlister, 1999] continuous zoom).
- Definitions/metadata for headings and cells are provided in pop-ups (tool-tip) as users mouse over headings or cells.
- The rows/columns/cells can be selected for closer display, creating a subtable, saving, printing, etc.

- Cells can contain multimedia objects as well as numeric values.

We are exploring ways to make statistical objects (e.g., cells) exhibit behavior. For example, the slider tool in one column (satisfaction) illustrates the possibility of users setting ranges on data to control display attributes or subsequent manipulation parameters. Likewise, some cells have special functionality—e.g., the comments column is editable; likewise, blood type is password protected and clicking on it and entering a password allows one to select/edit a blood type.

Figure 4. Table Browser Prototype Interface



The hierarchical list window at lower left contextualizes the user within all government data. A different version makes this a floating vertical window that may be more familiar to people used to Windows and WWW browser layouts. Drilling down through the hierarchical tables gives the users some sense of the mass of government tables and also contextualizes the table that is currently displayed. The lower right window is for URLs or other links. Note that the table in Figure 4 uses fictitious data to illustrate a range of features and we use actual tables from our partner agencies in next iterations.

Some manipulation features are already implemented and we are working to provide at least the following: display columns of numeric values by percentages; sort columns by various attributes appropriate to the data (e.g., rank, date, alphabetic), highlight maximum and minimum values in columns; show variances or confidence intervals for columns or cells respectively; and provide simple statistical summaries such as means. Good metadata is essential to support these operations, to give helpful elaborations of values upon mouse-over, and to build in "suggestions" to users when inappropriate operations are requested (e.g., means of categorical data, aggregating data with different base measures).

SUMMARY

Our research project aims to broaden access and use of the rich sets of statistical information collected and disseminated by the US Federal government in tabular formats. We believe that our work will have broader application to other data providers internationally. Our approach is to build sets of alternative interfaces that are rooted in studies of user behavior and needs. The user studies help define grammars for actions and queries, which in turn are mapped to data characteristics. We are using this principled approach to build interfaces for exploring data (overview and preview) with dynamic queries, posing and clarifying text-based queries, and understanding and using tabular data. These interfaces depend on good metadata techniques to provide explanations on demand.

There are many challenges to be met as we continue to develop and test these interfaces and as government electronic services improve.

- Extracting metadata and automating XML markup would facilitate retrieval as well as amendments and additions.
- Providing broad ranges of citizens with help and support will require more than cascading sets of on-demand elaborations and explanations to supplement today's email and phone help services with online communities.
- Balancing advanced interfaces with the installed base of hardware and software in the entire population requires good policy to adopt and evolve standards.
- Optimizing data flow from server to client to enable dynamic query techniques requires good data structuring as well as good bandwidth.
- Working with an array of government agencies with well-established and agency-specific work cultures requires good coordination and cooperation among agency leaders and policy makers.
- Expanding the range of interfaces and services to the broadest range of citizens will require both political will as well as technical advances.

Ultimately, it is crucial that such challenges be met if we are to continue to make government information more accessible and useful to the populace. This is a necessary condition for making government more responsive to citizen needs.

ACKNOWLEDGMENTS

This work is supported by US National Science Foundation Digital Government Initiative Grant #9876640. The authors acknowledge the work of Ben Brunk,, Fred Gey, Xiaming Mu, and Egemen Tanin on various aspects of this research.

REFERENCES

1. Bederson, B. & McAlister, B. (1999) Jazz: An Extensible 2D+ Zooming Graphics Toolkit in Java CS-TR-4015, UMIACS-TR-99-24, May 1999.
2. Brinton, W. (1914). Graphic methods for presenting facts. NY: The Engineering Magazine Co.
3. Carr, D. B. 1998. Multivariate Graphics, Encyclopedia of Biostatistics, Eds. P. Armitage and T. Colton, Vol. 4, pp. 2864-2886
4. Doan, K., Plaisant, C., Shneiderman, B. and Bruns, T. (1997) Interface and data architecture for query previews in networked information systems *ACM Transactions on Information Systems,* July 1999, Vol. 17, No. 3, 320-341
5. Greene, S., Marchionini, G., Plaisant, C., & Shneiderman, B. (2000). Previews and overviews in digital libraries: Designing surrogates to support visual information seeking. *Journal of the American Society for Information Science, 51*(4), 380-393.
6. Hall. R. (1943). Handbook of tabular presentation: How to design and edit statistical tables, a style manual and case book. NY: The Ronald Press Co.
7. Hert, C. & Marchionini, G. (1998). Information seeking behavior on statistical websites: Theoretical and design implications. *Proceedings of the American Society for Information Science Annual Meeting* (Pittsburgh, PA, oct. 25-29, 1998). 303-314.
8. Hyland, P. & Gould, T. (1998). External statistical data: Understanding users and improving access. International Journal of Human-Computer Interaction, 10(1), 71-83.
9. Liddy, E.D. (1998). Beyond Retrieval. In Martha Williams, ed. *Proceedings of the 19th Annual National Online Meeting* (New York). Medford, NJ: Learned Information. pp. 229-233.
10. Livny, M., Ramakrishnan, R., Beyer, K., Chen, G., Donjerkovic, D., Lawande, S., Myllymaki, J., & Wenger, K.DEVise: Integrated Querying and Visual Exploration of Large Datasets. (1997). Proceedings of ACM SIGMOD, May, 1997.
11. Moore, D.S. (1997). New pedagogy and new content: The Case of statistics. *International Statistical Review. 65* (2):123-165.
12. Pyreddy, P. & Croft, B. (1997). TINTIN: A system for retrieval in text tables. Proceedings of ACM Digital Libraries '97 (Philadelphia, July 23-26, 1997), 193-200.
13. Rao, R. & Card, S. (1994). The table lens: Merging graphical and symbolic representations in an interactive focus+context visualization for tabular information. Proceedings. of ACM CHI 94 (Boston, MA) NY: ACM Press. 318-322.
14. Shneiderman, B. (1998), *Designing the User Interface: Strategies for Effective Human-Computer Interaction (3rd Ed.).* Reading, MA: Addison Wesley Longman.
15. Tanin, E., Plaisant, C., & Shneiderman, B. (2000). Generalizing Query Previews**:** Broadening Access to Large Online Databases. Submitted to CUU 2000.
16. Tufte, E. R. (1983). Visual display of quantitative information. Cheshire, Conn.: Graphics Press.
17. Walker, H. & Durost, W. (1936). Statistical tables: Their structure and use. NY: Bureau of Publications Teachers College, Columbia University.
18. Wainer, H. (1997). Visual revelations: Graphical tales of fate and deception from Napolean Bonaparte to Ross Perot. NY: Copernicus Books.
19. Wilkinson, L. (1999). *The Grammar of Graphics*. New York: Springer.