Supporting Statistical Electronic Table Usage by Citizens

By Carol A. Hert, Elizabeth D. Liddy, Ben Shneiderman, and Gary Marchionini

Over 70 agencies at the federal level are charged with collecting data and producing and disseminating statistics. These statistics are used to inform government policy, shape health care initiatives, provide information on the state of the economy, and others. They also have significant impact on the lives of citizens who use the statistics, for example, to determine job opportunities, changes in social security benefits, and quality of life in particular areas. Our digital government project developed several specific technologies to support the location, manipulation, and understanding of a quintessential format for statistical information—the table.

Information seekers tend to employ two major strategies for location: an analytic or query formation strategy and a browsing strategy, depending on their information needs, personal characteristics, and the system [2]. Thus the first component of our work is in the provision of multiple location techniques: a natural language processing (NLP) environment and a browsing/exploration environment (the Exploratory Overview technique), to support statistical information seeking. It is also possible a user might retrieve both pre-made summary tables and the very raw datasets from which such tables would be built (by choosing specific variables, specific values of variables, and so on). The Exploratory Overview technique enables browsing these datasets. Once a user identifies the table or tables of interest, the user is likely to want to display a table and begin to work with it. The Table Browser supports a variety of manipulations and explanation functions to support this challenge. Underlying these three technical components is a rich knowledge of statistical information seeking behavior and the barriers users experience as they work with statistical tables.

Getting people to the right table or set of tables is a complex query interpretation task. The team built a statistical-query sublanguage grammar using NLP techniques [1]. This grammar will enable the system to automatically recognize predictable aspects of statistical queries and map them into the pre-established statistical query frames that, in turn, will be matched against the metadata describing the content of each table. The NLP capabilities enable accurate,
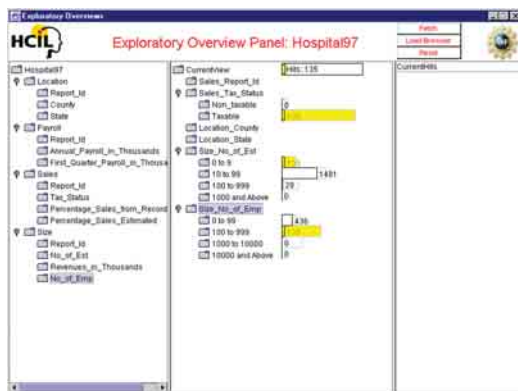
Figure 1. The data distribution information is attached to the buckets of three attributes expanded in the user view and multiple selections are made on these buckets.

Figure 2. (a) depicts the basic TB interface with a hierarchical list of tables in the upper-left pane, an extended metadata pane in the lower left, and the main table with a tool-tip explanation in the main pane. (b) shows four tables juxtaposed for comparison.

efficient mapping from the elements and relationships expressed in a user's statement of their information need to a table or tables that have the potential for addressing that information need.

As data volumes grow, the potential increases for user frustration, wasted network capacity, and increased server loads. We believe effective overviews and previews of databases and specific data sets can simultaneously improve the user experience and lighten system loads. The Exploratory Overview Panel allows users to see the distribution of data visually with histograms, maps, or textual lists prior to making choices and retrieving data. For example, in Figure 1, the panel enables users to make queries incrementally and visually by selecting items from a set of bar charts. Users get continuous feedback on the data distribution and result set size as they continue their selections, thereby avoiding wasted time on zero-hit or mega-hit queries.

Our third technology concerns the representations of data that a user can access and/or create on the fly. The Table Browser tool ([3], Figure 2) provides basic tabular functionality (for example, "sticky" headings that do not scroll with data, data reorganization capabilities), the ability to

retrieve explanatory metadata (via pop-ups, mouseovers, among others), all while minimizing the perceptual and cognitive effort of users. The tool underwent a series of usability studies and two eye-tracking studies [2]. The Java applet prototype reads XML files (of tabular data and associated metadata). It is available as an open source package from www.ils.unc.edu/idl/projects.html#stats.

This work has also provided insights into how electronic tables
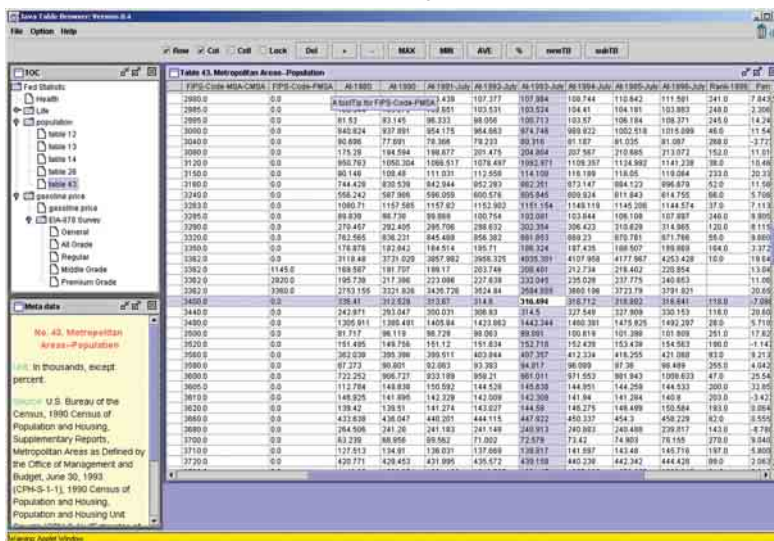
(a)
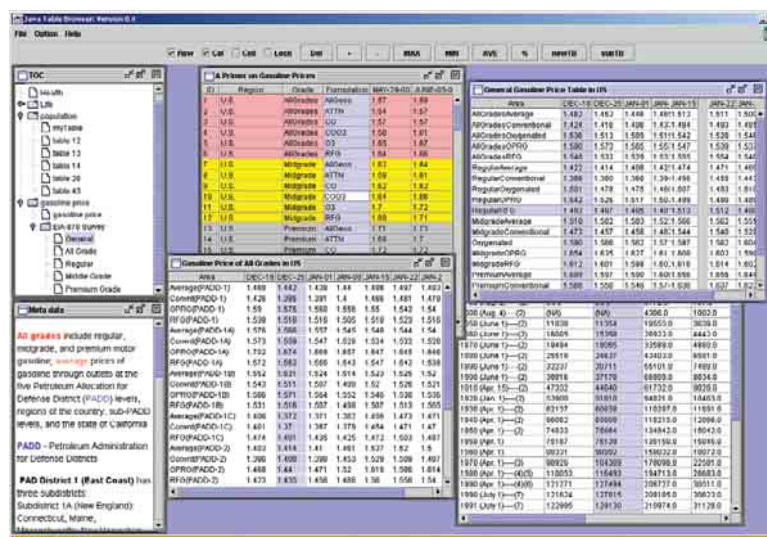
(b)

(e-tables) can be different from the traditionally static tabular representation in responding to a given user's particular knowledge and needs. Users will be able to generate the tables that match their needs exactly rather than retrieving a pre-made table and deriving appropriate information from it. Within a given table, they might also be able to perform certain arithmetic, sorting, and comparison operations easily. The notion of a set of tables that a user might need to juxtapose in his or her own physical or virtual world is likely to be replaced by the completely customized table built from rawer components. In the electronic environment, an e-table and its contents (cells, rows, columns, and so on) will be linked to the metadata that provides context and explanation thus enabling appropriate and informed usage and enabling users to learn as they go. **c**

REFERENCES

1. Liddy, E.D. and Liddy, J.H. An NLP approach for improving access to statistical information for the masses. In *Proceedings of the Federal Committee on Statistical Methodology 2001 Research Conference* (Washington, D.C., Nov. 14, 2001); www.fcsm. gov/01_papers/Liddy.pdf
2. Marchionini, G., Hert, C., Shneiderman, B., and Liddy, L. E-tables: Non-specialist use and understanding of statistical data. In *Proceedings of National Conference for Digital Government*. (Los Angeles, May 21–23, 2001), 114–119
3. Mu, X. and Marchionini. G. An architecture and prototype interface for an online statistical table browser. In *Proceedings of the Annual Meeting of the American Society for Information Science* (Washington, D.C., Nov. 5–8, 2001), 156–170.
4. Tanin, E. and Shneiderman, B. *Exploration of Large Online Data Tables Using Generalized Query Previews*. University of Maryland Computer Science Technical Report (June 2001).

CAROL A. HERT (cahert@syr.edu) is the director of and a professor in the School of Information Studies at Syracuse University, Syracuse, NY.
ELIZABETH D. LIDDY (liddy@syr.edu) is the director and a professor in the Center for Natural Language Processing in the School of Information Studies at Syracuse University, Syracuse, NY.
BEN SHNEIDERMAN (ben@cs.umd.edu) is a professor in the Department of Computer Science at the University of Maryland, and the founding director of the Human-Computer Interaction Lab.
GARY MARCHIONINI (march@ils.unc.edu) is the Cary C. Boshamer Distinguished Professor in the School of Information and Library Science at the University of North Carolina, Chapel Hill.