BIBL NEEDS REVISION

FCSM Statistical Policy Seminar
Session Title: **Integrating Electronic Systems for Disseminating Statistics**
Paper Title: **Interacting with Tabular Data Through the World Wide Web**

Paper Authors: **Carol A. Hert, Syracuse University; Gary Marchionini, Univ. of North Carolina; Elizabeth D. Liddy, Syracuse University; Ben Shneiderman, University of Maryland**
Paper Presented by: **Carol A. Hert** hert_c@bls.gov; **202-691-7399**

INTRODUCTION

People in all walks of life are often faced with a decision or want information that involves statistics. They may be considering moving to a location that would provide better job prospects or want to understand the quality of health care in their region. Perhaps they are considering getting an advanced degree in a scientific field and are concerned about job prospects in universities. Finding information on these topics often involves accessing statistics, frequently in the form of tables. Critical issues for these people, and for the agencies that produce and disseminate such information, is whether the users find tables that address their need(s) and having found them, whether the users are able to understand and use the information contained in those tables. We are interested in understanding what is possible with electronic tables and how people can best use them to meet their needs. We are exploring this domain and developing prototype interfaces that enable "the person on the street" to find and use statistical information presented tabularly[1].

In a world increasingly dominated by non-textual data and multi-media information provided via the World Wide Web, addressing these challenges will be essential for assuring that the progress made to date in access to textual information continues[2].

(INSERT FIGURE 1 HERE)

INTEGRATION ISSUES AND APPROACHES

Our work has involved integration of an understanding of user behavior[3] with the technical aspects of information provision and specification of a set of relationships among these aspects that will drive our ongoing work. This approach addresses several important issues currently dominating discussions within parts of the statistical community. The first two of these reflect a concern with how statistical agencies can serve an increasingly diverse and oftentimes, non-knowledgeable set of users. Tools and

[1] Our work is funded by the National Science Foundation and by the Bureaus of Labor Statistics and the Census.

[2] Visualizations, such as maps, scattergrams, or network diagrams are also helpful for user understanding, but are not the focus of this NSF-supported project.

[3] Findings In this area are not detailed here but are available In Hert and Marchionini (1997); Marchionini (1998, 1999) and Hert (1998, 1999).

techniques that worked for experts who knew the data and how to access them are not appropriate for less expert users. Novice users are less likely to understand particular statistical agencies, surveys, data sets, concepts, variables, and associated terminology.

The first issue concerns the ways in which people not well-versed in the domain express their information needs. Use of search engines (and the shift to natural language processing search engines) requires a user to express his or her information need. Research (Haas and Hert, 2000) demonstrates a limited overlap among terms an agency might use to describe a concept and those used in search engines. The nature of the mismatch includes terms not used at all by users (but important in agency language) and vice versa, and mismatches in specificity. Long standing evidence in information science also demonstrates that users are not able to express what it is they do not know. Getting people to the right table or set of tables is thus a complex interpretation task, one that Liddy, in the context of this project, is investigating. Liddy, drawing on her expertise in natural language processing systems, is drawing on user information need expressions (from email messages) to build a statistical sub-grammar. This grammar, used in conjunction with other natural language processing tools, will enable mapping from elements and relationships present in a user's expression of an information need to a table or tables that have potential to address that information need.

To better understand the dimensions of interest in typical queries, as well as the linguistic regularities that can be captured in a statistical-query sublanguage grammar which, in turn, will produce an internal query representation which maps the user's query dimensions into the tables' metadata, the team conducted the following tasks:

- Analyzed 1,000 actual user queries gathered from logs of users' seeking statistical information
- Developed an ontology of query dimensions using a data-up analysis based on the queries, and then extended the ontology where necessary with values from tables
- Produced a first draft of a statistical-query sublanguage grammar which will now be validated on a set of new queries

The analysis of the email queries showed that these users typically are concerned with a certain population, a quantification, a location, a time period, and a condition. The sublanguage grammar recognizes these elements by utilizing various levels of Natural Language Processing and maps them into the ontology we developed which captures the who, what, where, and when of statistical information queries. Our work is now focused on mapping the representation produced by the sublanguage grammar with the identifying text found in tables (e.g. table name, row, column), as well as the metadata accompanying the tables. Future work will focus on utilizing these capabilities in a retrieval engine that will provide the answer to the user's query, presented in such a way that 'every cell tells a story' so that the user will understand the full meaning of the specific statistical information they retrieve.

Along with terminological difficulties, our primary populations of interest have limited knowledge of the nuances of the agencies and the information provided. For example, many people are interested in the cost of living but probably are not aware that the Consumer Price Index, while providing insight into the cost of living, is not a direct correlate, nor are they likely to know which items are in the "market basket" and how those items change in different geographic areas or over time. Given this lack of knowledge, metadata and their representation become increasingly critical for providing explanation and the context surrounding a number or a set of numbers. Metadata initiatives within the statistical domain are in their and this project is serving to clarify which components of metadata are relevant to user understanding, how they might be conveyed, etc. Hert and Marchionini are taking the lead on this component of the project. Using specific examples of tables provided by our agency partners[4], the team is identifying:

- questions or uncertainties users might have related to the tables (to date we have worked with seniors and now are working with university students and faculty)
- answers to these questions (the metadata) and their location (e.g. in a web document, paper document, in a human's head)
- potentially relevant elements in the statistical metadata sets being developed by the Document Definition Initiative (DDI; URL: http://www.icpsr.umich.edu/DDI) and by the International Standards Organization, ISO; ISO/IEC11179 with statistical metadata extension)

The resultant knowledge of available metadata, their location, and how they correspond to user inquiries is an important input into the design of interface tools discussed below.

Along with facilitating search tasks and table understanding, we are also concerned with the potential of "information overload" as available information increases. As data volumes grow, the potential increases for user frustration, wasted network capacity, and increased server loads. We believe that effective overviews and previews of databases and specific data sets can simultaneously improve the user experience and lighten system loads. Well-designed search techniques and fill-in the-blank templates are sometimes effective for knowledgeable users; menu selection and visual queries can be highly effective for most users and many common tasks, thus one component of our project, under the direction of Shneiderman, is investigating such tools for exploration.

Our query preview prototypes have been tested with users and refined to make them still more effective. Instead of asking users to type in a date range, state name or variable name only to find that no such data is available, users can see the distribution of data visually with histograms, maps, or textual lists. For example in Figure 2, the Exploratory Overview 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

[4] The United States Bureaus of Labor Statistics, Census, and Justice Statistics, the National Center for Health Statistics, the Energy Information Administration, the National Science Foundation Science, and the Environmental Protection Agency

or mega-hit queries. We continue to refine our Java-based implementation to broaden its applicability and functionality.

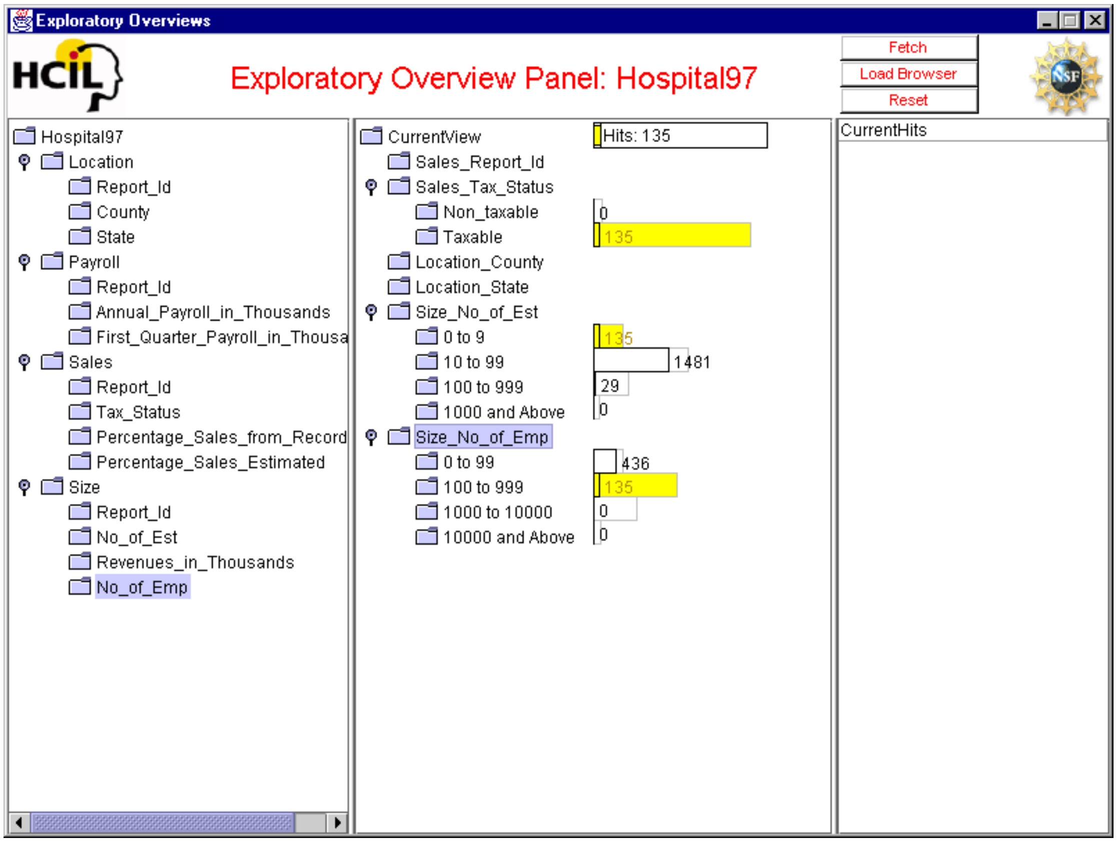


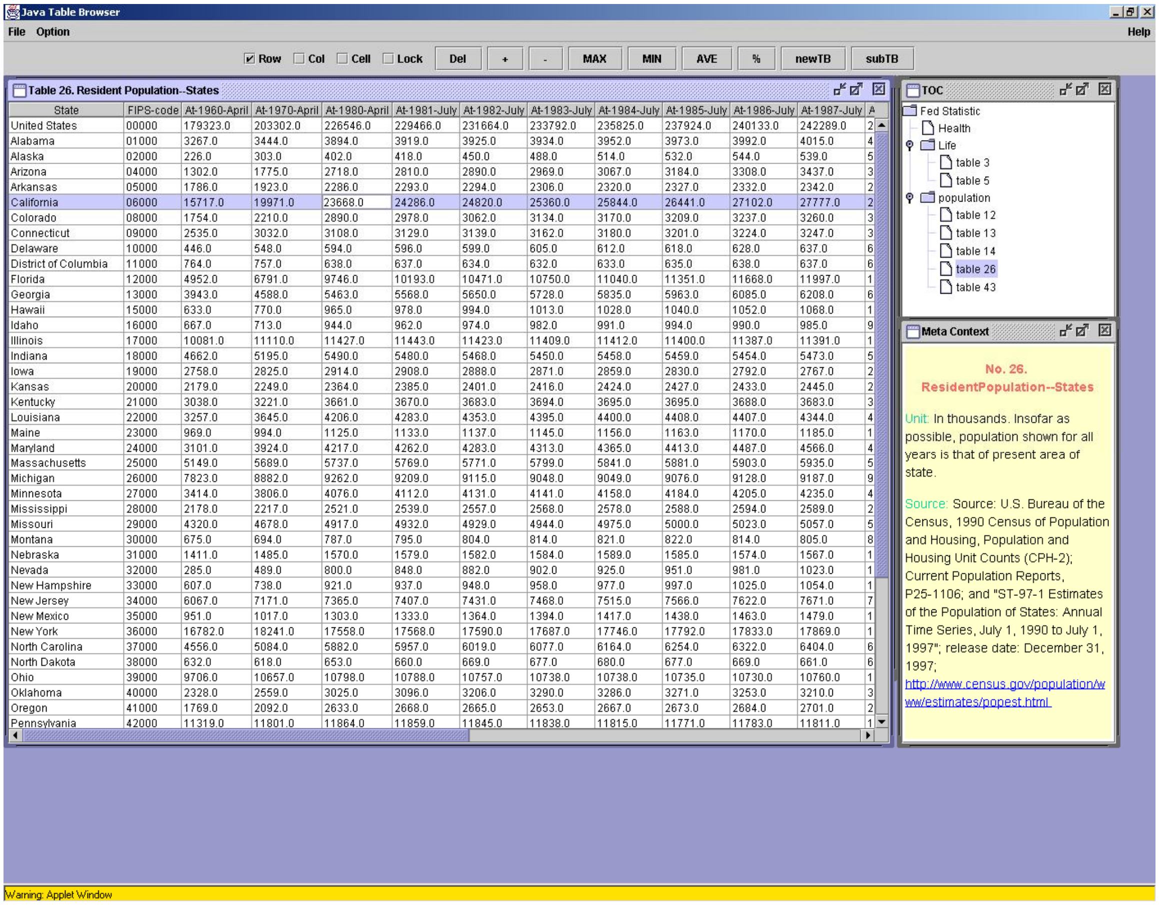
Figure 2: The data distribution information is attached to the buckets of three attributes expanded in the user view and multiple selections are made on these buckets

A third concern is with the representations of data that a user can access and/or create on the fly. Tables are a ubiquitous representational format both for storage and presentation. While there is a large literature on how to construct tables in paper formats, there is less knowledge on how to represent tables in electronic format for searching, exploration, browsing, and sharing, and how to link metadata to those tables. Producers of statistical data are also concerned with appropriate usage of their data and would like to find strategies to minimize incorrect comparisons, etc. We are taking advantage of recent developments that use XML (the DDI) and international standardization efforts (ISO/IEC 11179) in these investigations.

People using E-tables should be able to leverage the dynamic capabilities of the computer to display and manipulate data easily and according to their specific needs. Our preliminary prototype supports many of the tasks we have identified in our user studies while eliminating many of the limitations typically placed on web-based tables. (See Figure 3). It 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
- A simple zoom in and out is provided
- Definitions/metadata for headings and cells are provided in pop-ups (tool-tip) as users mouse over headings or cells.
- The metadata for the overall table shows up in a window (lower right corner in the figure). This window can be resized and may include hyperlinks to glossaries or related tables.
- 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.
- The specific table is contextualized in a hierarchical list of related tables (shown in the window upper right in the figure).

The prototype is implemented as a Java applet that reads XML files with the data and renders the table according to client preferences (e.g., browser, display, etc.). We are currently marking up the XML files manually in order to develop a project document type definition (DTD) that will drive automatic markup from the underlying databases containing the data.



State	FIPS-code	At-1960-April	At-1970-April	At-1980-April	At-1981-July	At-1982-July	At-1983-July	At-1984-July	At-1985-July	At-1986-July	At-1987-July	A
United States	00000	179323.0	203302.0	226546.0	229466.0	231664.0	233792.0	235825.0	237924.0	240133.0	242289.0	2
Alabama	01000	3267.0	3444.0	3894.0	3919.0	3925.0	3934.0	3952.0	3973.0	3992.0	4015.0	4
Alaska	02000	226.0	303.0	402.0	418.0	450.0	488.0	514.0	532.0	544.0	539.0	5
Arizona	04000	1302.0	1775.0	2718.0	2810.0	2890.0	2969.0	3067.0	3184.0	3308.0	3437.0	3
Arkansas	05000	1786.0	1923.0	2286.0	2293.0	2294.0	2306.0	2320.0	2327.0	2332.0	2342.0	2
California	06000	15717.0	19971.0	23668.0	24286.0	24820.0	25360.0	25844.0	26441.0	27102.0	27777.0	2
Colorado	08000	1754.0	2210.0	2890.0	2978.0	3062.0	3134.0	3170.0	3209.0	3237.0	3260.0	3
Connecticut	09000	2535.0	3032.0	3108.0	3129.0	3139.0	3162.0	3180.0	3201.0	3224.0	3247.0	3
Delaware	10000	446.0	548.0	594.0	596.0	599.0	605.0	612.0	618.0	628.0	637.0	6
District of Columbia	11000	764.0	757.0	638.0	637.0	634.0	632.0	633.0	635.0	638.0	637.0	6
Florida	12000	4952.0	6791.0	9746.0	10193.0	10471.0	10750.0	11040.0	11351.0	11668.0	11997.0	1
Georgia	13000	3943.0	4588.0	5463.0	5568.0	5650.0	5728.0	5835.0	5963.0	6085.0	6208.0	6
Hawaii	15000	633.0	770.0	965.0	978.0	994.0	1013.0	1028.0	1040.0	1052.0	1068.0	1
Idaho	16000	667.0	713.0	944.0	962.0	974.0	982.0	991.0	994.0	990.0	985.0	9
Illinois	17000	10081.0	11110.0	11427.0	11443.0	11423.0	11409.0	11412.0	11400.0	11387.0	11391.0	1
Indiana	18000	4662.0	5195.0	5490.0	5480.0	5468.0	5450.0	5458.0	5459.0	5454.0	5473.0	5
Iowa	19000	2758.0	2825.0	2914.0	2908.0	2888.0	2871.0	2859.0	2830.0	2792.0	2767.0	2
Kansas	20000	2179.0	2249.0	2364.0	2385.0	2401.0	2416.0	2424.0	2427.0	2433.0	2445.0	2
Kentucky	21000	3038.0	3221.0	3661.0	3670.0	3683.0	3694.0	3695.0	3695.0	3688.0	3683.0	3
Louisiana	22000	3257.0	3645.0	4206.0	4283.0	4353.0	4395.0	4400.0	4408.0	4407.0	4344.0	4
Maine	23000	969.0	994.0	1125.0	1133.0	1137.0	1145.0	1156.0	1163.0	1170.0	1185.0	1
Maryland	24000	3101.0	3924.0	4217.0	4262.0	4283.0	4313.0	4365.0	4413.0	4487.0	4566.0	4
Massachusetts	25000	5149.0	5689.0	5737.0	5769.0	5771.0	5799.0	5841.0	5881.0	5903.0	5935.0	5
Michigan	26000	7823.0	8882.0	9262.0	9209.0	9115.0	9048.0	9049.0	9076.0	9128.0	9187.0	9
Minnesota	27000	3414.0	3806.0	4076.0	4112.0	4131.0	4141.0	4158.0	4184.0	4205.0	4235.0	4
Mississippi	28000	2178.0	2217.0	2521.0	2539.0	2557.0	2568.0	2578.0	2588.0	2594.0	2589.0	2
Missouri	29000	4320.0	4678.0	4917.0	4932.0	4929.0	4944.0	4975.0	5000.0	5023.0	5057.0	5
Montana	30000	675.0	694.0	787.0	795.0	804.0	814.0	821.0	822.0	814.0	805.0	8
Nebraska	31000	1411.0	1485.0	1570.0	1579.0	1582.0	1584.0	1589.0	1585.0	1574.0	1567.0	1
Nevada	32000	285.0	489.0	800.0	848.0	882.0	902.0	925.0	951.0	981.0	1023.0	1
New Hampshire	33000	607.0	738.0	921.0	937.0	948.0	958.0	977.0	997.0	1025.0	1054.0	1
New Jersey	34000	6067.0	7171.0	7365.0	7407.0	7431.0	7468.0	7515.0	7566.0	7622.0	7671.0	7
New Mexico	35000	951.0	1017.0	1303.0	1333.0	1364.0	1394.0	1417.0	1438.0	1463.0	1479.0	1
New York	36000	16782.0	18241.0	17558.0	17568.0	17590.0	17687.0	17746.0	17792.0	17833.0	17869.0	1
North Carolina	37000	4556.0	5084.0	5882.0	5957.0	6019.0	6077.0	6164.0	6254.0	6322.0	6404.0	6
North Dakota	38000	632.0	618.0	653.0	660.0	669.0	677.0	680.0	677.0	669.0	661.0	6
Ohio	39000	9706.0	10657.0	10798.0	10788.0	10757.0	10738.0	10738.0	10735.0	10730.0	10760.0	1
Oklahoma	40000	2328.0	2559.0	3025.0	3096.0	3206.0	3290.0	3286.0	3271.0	3253.0	3210.0	3
Oregon	41000	1769.0	2092.0	2633.0	2668.0	2665.0	2653.0	2667.0	2673.0	2684.0	2701.0	2
Pennsylvania	42000	11319.0	11801.0	11864.0	11859.0	11845.0	11838.0	11815.0	11771.0	11783.0	11811.0	1

No. 26.
ResidentPopulation--States

Unit: In thousands. Insofar as possible, population shown for all years is that of present area of state.

Source: Source: U.S. Bureau of the Census, 1990 Census of Population and Housing, Population and Housing Unit Counts (CPH-2); Current Population Reports, P25-1106; and "ST-97-1 Estimates of the Population of States: Annual Time Series, July 1, 1990 to July 1, 1997"; release date: December 31, 1997; http://www.census.gov/population/www/estimates/popest.html

Warning: Applet Window

Figure 3: The Table Browser

INTEGRATION AND ONGOING WORK

This project integrates all these aspects related to finding and using statistical tables together as represented in Figure 1. The three interfaces/system components are represented as the three circular entities: the natural language processor with a statistical-query sublanguage grammar to interpret user queries, the data set exploration tool that enables users to understand data sets prior to downloading them, and the table browser tool that is used to render tables retrieved via both tools. These three tools are embedded in the larger knowledge of user information needs (reported in Hert and Marchionini, 1997), user query structure (developed in Hert and Marchionini and undergoing expansion in this project), user interaction with the exploration and browsing tools (from usability tests within this project), and knowledge of metadata and their utility to users (outlined previously in this paper)

Our work for this year involves expanding the set of tables that our prototype interface tools can process, ongoing work with users to test the interfaces, and the generalization of our work so that it can support the rich range of tables that are produced (or could be produced) by the statistical agencies.

REFERENCES

Hert, C.A. (1999). Federal Statistical Website Users And Their Tasks: Investigations Of Avenues To Facilitate Access: Final Report to the United States Bureau of Labor Statistics. Available at: http://istweb.syr.edu/~hert/BLSphase3.PDF

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

Haas, S.W. and Hert, C.A. (2000). Terminology Development and Organization in Multi-Community Environments: The Case of Statistical Information. In Soergel, D. (ed.) Proc. 11th ASIS&T SIG/CR Classification Research Workshop. Chicago, IL, November 12, 2000, p. 51-72

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

Marchionini, G. (1998) Advanced Interface Designs for the BLS Website: Final Report to the Bureau of Labor Statistics. http://ils.unc.edu/~march/blsreport98/final_report.html

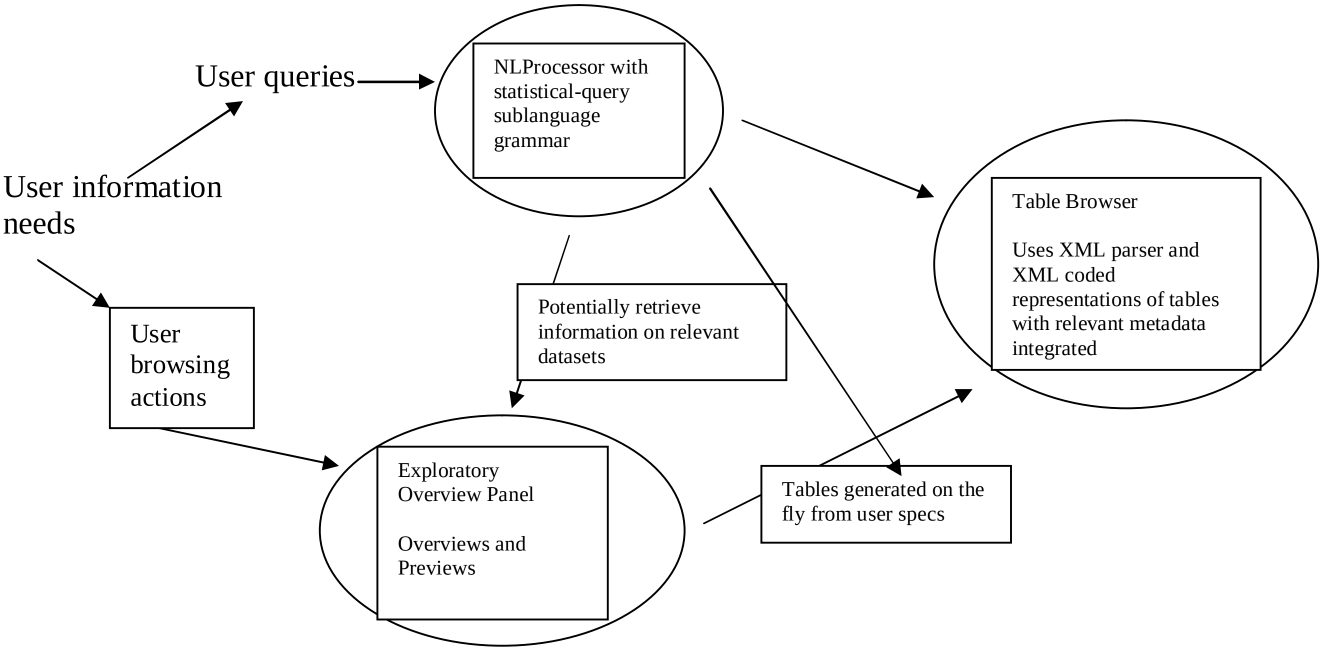
Marchionini, G. (1999). An Alternative Site Map Tool for The FedStats Statistical Website. http://ils.unc.edu/~march/bls_final_report99.pdf

Marchionini, G. (2000). From Overviews to Previews to Answers: Integrated Interfaces for Federal Statistics Report to the Bureau of Labor Statistics, June 30, 2000 http://ils.unc.edu/~march/bls_final_report_99-00.pdf

Greene, S., Tanin, E., Plaisant, C., Shneiderman, B., Olsen, L., Major, G., and Johns, S., The end of zero-hit queries: Query previews for NASA's Global Change Master Directory, International Journal of Digital Libraries 2, No.2+3 (1999), 79-90.

Tanin, E., Beigel, R., and Shneiderman, B., 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).

FIGURE 1: INTEGRATION


User information needs
User queries
NLProcessor with statistical-query sublanguage grammar
User browsing actions
Potentially retrieve information on relevant datasets
Table Browser
Uses XML parser and XML coded representations of tables with relevant metadata integrated
Exploratory Overview Panel
Overviews and Previews
Tables generated on the fly from user specs