

# Interacting with Statistics

## Report from a Workshop at CHI 99

Michael Levi and Frederick Conrad

### Overview

Suppose you are moving to a new city. You might reasonably want to know if you should worry about crime. Or whether the local schools are good. Whether the drinking water is clean. And whether the job prospects are encouraging. In each of these cases the relevant information exists in one or more statistical data bases (SDBs). Whether a typical information seeker can find the information, however, is another question.

Information seekers are granted access to the content of statistical databases through user interfaces, which act as gatekeepers. The quality of the interface – its suitability for a particular user population performing a particular set of tasks – is the prime determinant of successful accessibility.

The importance of usability engineering to SDBs has been brought into sharper focus since the World Wide Web has become a popular medium for disseminating statistics. The usability issues are essentially the same, however, whether access to statistical data takes place over the Internet, over a LAN, or from a CD drive on the user's desktop. The authors organized a workshop at the CHI 99 Conference on Human Factors in Computing Systems to explore these usability issues with developers of SDB interfaces and HCI practitioners who have worked on the problems of retrieving statistical data.

The purpose of the workshop was to foster the development of better, more usable interfaces to statistical databases, by which we mean large, numeric data sets concerned primarily

with social science issues, which may be directly accessible by the public. More concretely, the workshop concentrated on the important HCI problems that are peculiar to:

- 1) Selection, i.e. communicating requests for statistical information, and
- 2) Display, i.e. presenting the results of those requests in a form the users can interpret.

The workshop's progress and insight came from the interdisciplinary discussion and collaboration between the participants. The idea was for SDB interface developers to expand their working knowledge of user-centered design and for HCI practitioners to sharpen their appreciation of the design challenges in this domain.

### Participants

John Bosley (Bureau of Labor Statistics)  
Cavan Capps (Bureau of the Census)  
Dan Carr (George Mason University)  
Kathleen Donohue (Bureau of Labor Statistics)  
Richard Esposito (Bureau of Labor Statistics)  
Lee H. Giesbrecht (Bureau of Transportation Statistics)  
Per Henrik Johansen (Statistics Norway)  
Natalie N. LaPrade (SAS Institute)  
Gary Marchionini (University of North Carolina at Chapel Hill)

Frederick Conrad -- Organizer  
Michael Levi -- Organizer

## Summary of Position Papers

John Bosley and Cavin Capps describe a usability evaluation of a prototype tool for extracting and tabulating employment data from Current Population Survey databases. While the evaluation raised some doubt about the usability of certain features as implemented in this particular prototype, it confirmed the importance of following several design principles in building tools for extracting statistical data: (1) make it easy to use numerical data and metadata at the same time, (2) give users some kind of tool to help them rename and combine variables, and (3) give users flexibility in how they step through the data extraction process.

Daniel Carr argues that when statistical summaries are presented to the public their multiple facets should be made visually explicit. For example, in addition to showing the statistical estimates that are typically derived from a given data, data providers could graphically represent the geographic context and measures of both estimate quality and importance. Dan discusses linked micromaps (e.g. Carr, Olsen, Courbois, Pierson, & Carr, 1998), as one way to balance estimates, confidence bounds, related statistics, and spatial indices. Dan also suggests four classes of global quality indicators that should accompany plots and notes that studies across the federal agencies are not of equal quality.

Kate Donohue argues that statistical web sites need to be designed so that it is easy for users to obtain and correctly interpret data from multiple surveys and organizations, and therefore multiple databases. It is hard for users to work with data from multiple databases as well as to work with metadata (explanatory text about the data) from multiple sources. Thus users are prone to misinterpret the data and compare and combine them in unwarranted ways. Kate identifies five types of information that users frequently request which span multiple surveys and organizations: Geography, Industry, Occupation, Demographic Characteristics, and Subject/Topic. A short term solution, using geography as an example, would be to allow users to specify an area and then provide users with a clickable list of all data and metadata available for that area.

Rich Esposito argues that it is far easier to evaluate the quality of large data sets if the data are presented visually than numerically. Rich demonstrates this with several visual tools used by U.S. government analysts to detect and account for outliers (Esposito, Fox, Lin and Tideman). For example, the anomaly map is a circular tree diagram in which the nodes represent categories of industries from which establishments participate in a particular employment survey. If there has been abnormal activity in a particular industry in the last month, the corresponding node is colored to reflect this. The analyst explores the abnormality by clicking a colored node. This produces several displays including a plot of current month versus historical activity for individual establishments in the suspicious industry. Abnormal movement will be immediately evident if a point (a particular establishment) is substantially off the diagonal. By clicking these

points the analyst can inspect the numerical data for particular outlying establishments.

Lee Giesbrecht advocates making more information about data quality available to data users. For example, Lee would like to see information included data sets about the nonsampling error for each variable. This includes nonresponse rates, imputation rates, and estimates of bias, where they exist. Lee points to two prototype projects in the U.S. federal government that promise to produce sampling error estimates on the fly and may even add "error whiskers" to online graphs and charts. However, Lee calls for tools that do more to inform users about pitfalls in their analyses based on what is known about the quality of the data they are analyzing. For example, once the correct sampling error estimates for a data set are calculated, other tools could help users conduct significance testing and compute confidence intervals.

Per Henrik Johansen is concerned with designing interfaces to better disseminate statistical data from a central site to users from municipal governments – primarily "knowledgeable-intermittent users." He believes in making alternative interfaces available that focus on different facets of the data. The particular facets that Per Henrik has explored are region, time and topic. He discusses interfaces at Statistics Norway organized around each of these facets to: retrieve preformatted tables, navigate through a complex data base, and specify queries to a complex database. One lesson learned from these efforts is that the interface is important but not as important as the content. Thus far, users have asked more for additional data than additional means of access.

Natalie LaPrade describes the difficulty in creating graphical interfaces for end user analysis tools. Statistical packages such as SAS have traditionally allowed users to specify the exact parameters of an analysis through code or syntax. Natalie has discovered that there are not always clear graphical analogues to these syntactic methods. She has been designing an interface for the CONTRAST and ESTIMATE statements within the SAS Procedure "GLM."

Gary Marchionini describes a research program to study and develop interfaces to statistical tables. The approach enables users to interact with tables (e.g. manipulating columns) and to obtain explanatory text as needed (that is, when the user wants more information and when the system detects that the user needs more information for example to avoid unwarranted comparisons). Gary proposes developing an integrated table browser in which (1) row and column information is preserved as users scroll, (2) individual cells can be associated with detailed notes on units, acquisition, and usage, (3) scaling information is visually encoded (e.g. adjusting for uneven intervals and unit measures), (4) users can browse metadata as well as data tables, (5) multiple tables can be compared, and (6) users can zoom and view tables by particular properties.

## Workshop Agenda

### I. Orientation and Introductions

### II. Users

Who are our users? What tasks are they trying to accomplish? What is their level of literacy regarding statistics? Regarding computers? How motivated are they?

### III. Data

What characterizes the data we are using? To what are we building interfaces? Why is the data important? How large are the data sets? How complex are the relations between data items? Are there confidentiality issues? What are the other challenges, constraints?

### IV. Ease of Use vs. Statistical Richness

Is there a tension between making statistical data available to a wide audience and communicating its inherent complexity? How do we make creative use of this tension?

### V. Selection

What are the common usability problems associated with trying to locate a desired piece of information from a potentially very large (set of) tables(s) or statistical database?

Parallel discussion #1a, Site Structure: How can statistical Web sites be structured and relevant data pre-selected for optimal use?

Parallel discussion #1b, Queries: How can ad-hoc queries be implemented for optimal use?

### VI. Display

What are the common usability problems associated with displaying the results of a query against a statistical database?

Parallel discussion #2a, Numerical Data: How can numeric data (tables, database extracts) best be presented?

Parallel discussion #2b, Visualized Data: What visualization techniques (graphs, graphics) can most effectively be used to convey statistical information?

### VII. Conclusions

## Discussion at the Workshop

Several broad themes threaded their way through most of the workshop, seemingly relevant to most or all of the agenda topics.

1. Large statistical data sets are inherently complex, and this complexity must be adequately communicated to end users.

Statistical data providers tend to be acutely aware of both the richness and the limitations of the data they disseminate. Interrelationships between data items are often not straightforward, and frequently demand a quite sophisticated understanding of definitions, collection procedures, or statistical algorithms before valid comparisons can be made. For example, even though several different geographically-based surveys report data for "New York City", one of these surveys might actually include part of New Jersey in this item, another might include part of Connecticut, and a third might limit itself to the actual municipal boundaries. In a similar vein, directly comparing two "Laspeyres index" values (such as the Consumer Price Index for the United States and the Consumer Price Index for a particular area within the U.S.) is statistically meaningless. In this case only percent changes over identical time periods yield useful information.

At the same time, the top-level numbers, stripped of their context, are often widely quoted in lay publications, and frequently have immediate and significant economic or social impact. Thus interest in the numbers, and a desire to understand them better, is frequently high.

The trick for interface designers is to balance the presentation so as not to oversimplify, yet not to intimidate, either. Over the course of the workshop participants developed four underlying design principles to help us achieve this goal:

- Provide the explanatory information about the data (metadata) that is most relevant to end users.
- Inform users what they can and cannot legitimately do with data.
- Provide users a readily understandable indicator of quality.
- Use the interfaces to promote statistical literacy.

Discussions of data quality were intense enough to generate the following manifesto:

*The Web has made statistical data available like never before. Many new users lack the statistical background to judge the quality of this data. Since so many people make significant decisions based on these data, a readily understood indicator of statistical quality must be developed and, subsequently accompany every data set.*

Participants agreed to take this statement back to our home agencies and working groups and try to develop a broader constituency for a standard quality rating.

2. Users should be provided with alternative views of the data because there is no single best way to access or present data.

Users typically try to perform a wide variety of tasks when accessing statistical data sets. Even without comparing detailed site-specific and user-specific task analyses, some broad usage patterns emerged during discussions among participants:

- Zero in on a single value
- Compare a relatively small number of values
- Browse through a large set of values looking for patterns or trends

Each of these tasks demands a different interface. The workshop participants spent quite a bit of time explaining solutions they had tried – some successful, others less so – and brainstorming new approaches.

Solutions included standard cross-tabulations and other tabular presentations linked through a menu hierarchy or search engine; a variety of graphical depictions of data; and any number of ad-hoc database query interfaces.

One strategy of interest to almost all participants involved creating prepared “packets” of information on particular topics. This could be a pre-formatted HTML page organized around a specific geographical region, industry, mode of transportation, etc. It could also be a pre-defined partition of a database or any other compilation of data into subsets based on user interests.

Many participants were interested in maintaining user profiles; one idea was to implement referrer systems such as are increasingly common at vendor sites on the Web (“Many people who are interested in these unemployment statistics are also interested in the following employment figures...”)

3. There are numerous tensions in designing interfaces to statistical data bases.

Similar to virtually every other interface task the participants have been involved in, designing interfaces to statistical data sets involves a substantial number of trade-offs and compromises:

- Is a system intended for retrieving facts or analyzing data?
- What is important to users may not coincide with what data providers believe users should know.
- How does one differentiate between an interface which

is visually abundant and one that is simply cluttered?

- What is technically possible may not be organizationally or politically practical.
- How much weight should be placed on expert recommendations as opposed to participatory design.
- How might we balance some users’ expectations of relative permanence with the reality of ongoing data revisions and corrections.

Again similar to most of the other interface tasks we have been involved in, the solution to many of these tensions is “It depends.” Careful analysis of actual users and the tasks they are trying to accomplish is the most likely to yield useful results.

4. Data retrieval systems should be interactive.

On-line interfaces to statistical data sets should provide users with capabilities that can not be achieved on paper. For example,

- Interfaces need to be developed that preserve context while displaying detail.
- Footnotes should be displayed on demand, close to the item being annotated.
- Interfaces can simultaneously update multiple displays when a user takes an action in any one of the display areas.
- Users should be able to control which items are displayed, what aspects of the items are displayed, and what format the display takes.

Rather than be passive consumers of statistical data, users should be encouraged to do what the title of the workshop promises: interact with statistics.

### Informal Communications

In addition to the formal agenda, valuable informal discussions took place between participants. As a result of these conversations several future collaboration between Federal agencies and educational institutions are being planned, ranging from training efforts to joint development projects.

Ultimately there was a consensus among participants that the group had made real progress in addressing many of the problems – policy and technology-related – that face the designer of interfaces to statistical data sets. Of course more work remains to be done, but the workshop provided inspiration and motivation, as well as concrete ideas and approaches for the near-term future.

## About The Authors

Michael D. Levi is a project manager at the U.S. Bureau of Labor Statistics (BLS). He currently manages the BLS public access Web site, and was until recently responsible for software development, maintenance, and production in support of some of the largest employment and unemployment surveys conducted by the federal government.

Frederick G. Conrad is a cognitive psychologist in the Office of Survey Methods Research at BLS. His current work concerns developing and evaluating methods for collecting and disseminating statistical data. His non-HCI work often involves human-human interaction, for example how the interaction between survey interviewers and respondents affects the accuracy of the resulting data.

Michael Levi and Frederick Conrad are co-authors of several papers on usability testing and have spoken on the topic of usability testing and the World Wide Web to various government agencies and professional conferences. Mr. Levi and Mr. Conrad organized a workshop at CHI 97 on Usability Testing Web Sites, and organized and spoke on a panel at CHI 98 titled "Is the Web Really Different from Everything Else?"

## Authors' Addresses

Michael D. Levi  
Levi\_M@bls.gov  
2 Massachusetts Ave., NE  
Room 5110  
Washington, DC 20212, USA

Frederick G. Conrad  
Conrad\_F@bls.gov  
2 Massachusetts Ave., NE  
Room 4915  
Washington, DC 20212, USA

## References

Carr, D. B., Olsen, A.R., Courbois, J. P., Pierson, S. M. & Carr, D.A. (1998). "Linked Micromap Plots: Named and Described," *Statistical Computing & Graphics Newsletter*, 9 (1), pp. 24-32.

Esposito, R., Fox, J., Lin, D. & Tidemann (1994). ARIES: A visual path in the investigation of statistical data. *Journal of Computational and Graphical Statistics*, 3 (2), 113-125.

