Governmental Statistical Data on the Web: A case study of FedStats

Irina Ceaparu
University of Maryland, Department of Computer Science,
Human Computer Interaction Laboratory

Draft: October 5th, 2002

Introduction

Over 70 United States federal government agencies collect and store statistical data that eventually are made available to the general public. The World Wide Web offers the best medium for dissemination of these data. However, to facilitate access to these statistical data, a common portal with an easy to use interface is required. This portal is necessary to ensure that the general public, as well as researchers and statisticians, know about the existence of such data and can easily and quickly have access to the huge amount of information the federal agencies provide.

The only significant attempt to create such a portal began in 1997, when the FedStats web site (http://www.FedStats.gov) became publicly available. The site is designed to complement the already existing web sites of each federal agency, by providing a unique point of access to all collections of statistical data from almost 100 different agencies. The web site is intended to help users find the information they need without having to visit several web sites and without needing previous knowledge of the structure of the governmental agencies.

The objective of this paper is to analyze the FedStats web site and determine its usefulness to citizens. In this respect, a usability test was conducted, and the paper describes its results and the recommendations we make to the designers of the web site.

Previous Work and Related Literature

Bosley and Straub [1] have conducted multiple usability tests of statistical web sites, in which representative general public end users tried to find data to answer everyday questions that have statistical answers. Their analysis lead them to a set of 10 heuristics for interactive statistical database interfaces:

1. Give an overview of the available data	6. Use simple interaction schemes to accomplish complex query-building
2. Support situational awareness within the available data	7. Summarize outcome of complex data specifications for review and confirmation
3. Display and clearly define metadata	8. Offer choices of easy-to-interpret output formats
4. Put adequate and clear instructions on the interface	9. Design output formats to facilitate quick and reliable query validation
5. Link users to frequently requested analysis	10. Help users avoid searching for non-existent or non-available data

A team of researchers at the Digital Government Research Center conducted a project [2] about how to simplify data access. They built a system for disseminating statistical data from several federal agencies that consists of three parts: the database manager and access planner, the overarching ontology in which terms are defined and standardized, and the interface. The interface addresses the following issues: support for adaptive, context-sensitive queries via a system of guided menus; display of tables created by the integration back-end from one or multiple individual databases, along with footnotes and links to original data sources; and browsing of the ontology that supports the entire integration model, with the capability to display concept attributes, relationships and definitions in graphics and text.

The CHI 99 webshop "Interacting with Statistics" addressed several issues of designing interfaces to statistical databases.
Donohue [3] described two usability issues in analyzing the Bureau of Labor Statistics' web site: cross-survey data retrieval and metadata. In cross-survey data retrieval, in which searches are done across all relevant data sources based on criteria submitted by the user, the main challenge is the design of the interface. It must be intuitive, assist users in creating efficient queries, and return the output in a meaningful way. Donohue identifies five broad areas for cross-survey information: Geography, Industry, Occupation, Demographic Characteristics and Subject/Topic. The metadata, meaning the information about the data, impacts how the data can be used. The user must know how to identify the metadata and how to use it in order to check the correctness and usefulness of the answer.
Johansen [4], from the Division for IT Development at Statistics Norway, described the Kostra project - electronic data collection from municipalities to Statistics Norway and distribution of statistics from Statistics Norway back to the municipalities. He discussed three interfaces for accessing and interacting with regional statistics over the Internet:

Regions (a typical interface might be maps), Time (a time-line or scale would be an obvious approach for an implementation) and Variables (due to the nature of these, a list or tree structure could act as an interface).

The United Nations Statistical Commission and the Economic Commission for Europe proposed in 2000 a set of guidelines for statistical metadata on the Internet [5]. The guidelines refer to three types of metadata: metadata assisting search and navigation (e.g. sitemap/table of contents, descriptions of statistical subject areas and institutions, a list of key words linking everyday language to statistical tables/graphs, local search engine), metadata assisting interpretation (e.g. content description, measurement unit, time period, regional unit), and metadata assisting post-processing (e.g. for downloads, data/metadata should allow further processing using suitable tools, like spreadsheets, databases, packages for statistical analysis, etc.).

The United Nations Statistical Commission and the Economic Commission for Europe proposed in 2001 another set of guidelines for best practices in designing web sites for dissemination of statistics [6]. The following is a summary of the guidelines:
Users. They have to be the center of attention. It is important to investigate who the users are, what they really need, how they use the data, what their competence levels are. At least some kind of usability test should be organized, be it even with very simple tools. Usability studies provide an essential feedback that permits improvement of the web site and, in the long term, reduces costs necessary for redesign.
Maintenance. The development and maintenance of the web site of a statistical office should not just be the task of IT-experts and special dissemination staff. Strong support among the (top) management is one of the most essential success factors for a statistical web site. It must be ensured that the necessary competence to develop and maintain a web site is available.
Search and navigation. The web site architecture must provide comprehensible navigation across the whole web site. It must be easy to find the desired data inside the web site. The response time must have high priority.
Interpretation. A flexible and consistent metadata support should ensure that the published statistical data is transparent and comprehensible to the users; any kind of misunderstanding and misinterpretation should be avoided. The metadata should support the comparability of data over time, i.e. historical data should be supported by metadata. The published statistical data must be consistent across the whole web site.
Post-processing. The user must be able to download data into his/her own technical environment. The data should be provided in well-accepted standard formats.

In 1996, Gary Marchionini and Carol Hert [7]-[11] began working with the Bureau of Labor Statistics (BLS) and the FedStats group. Their goals were to gain a better understanding of the interaction of non-specialist users with statistical data and to analyze how federal statistical agencies can adopt and adapt technologies to better serve the needs of diverse constituencies. They developed a user task and type taxonomy, methodologies for assessing user behavior, user interface design guidelines, as well as a user interface tool called the Relation Browser for use in the FedStats web site.

In 1998 Stephanie Haas [12]-[14] joined the team and focused on understanding the linguistic mappings between end-user vocabularies and agency vocabularies. The Relation Browser tool was revised based on the usability tests and implemented as an alternative site map for FedStats and field tested over a nine-month period.

In 1999 Ben Shneiderman from the University of Maryland and Liz Liddy from Syracuse University [15]-[19] joined the team as well, focusing on making statistical tables from federal agencies easier to find and use. The academic team worked with several government agency partners (BLS, EIA, Census, NCHS) to study how people find and understand statistical tables. Metadata issues for statistical information were investigated and mapped to interface tools. Dynamic query tools and natural language processing techniques were developed to facilitate finding statistical tables and a table browser tool was developed and extensively usability tested.

In 2002 a new effort began as part of the NSF Digital Government Program [20] to study issues of integration requirements needed for a national statistical knowledge network. The goals are to work with FedStats agencies and state agencies to make government statistical data more accessible and understandable to non-specialists who need data to make better decisions; to address horizontal integration across federal agencies (e.g., interoperation for data and metadata vocabulary, interfaces, and the overall user experience), and vertical integration across levels of government (e.g., linking state and local data to federal sources).

In 2002, a project related to user frustrations with computers was started by the Human-Computer Interaction Laboratory in the Department of Computer Science at University of Maryland, the Department of Sociology at University of Maryland and the Department of Computer and Information Sciences & Center for Applied Information Technology at Towson University [21-23]. The project had 3 phases. The first study attempted to measure, through 111 subjects, the frequency, cause, and the level of severity of frustrating experiences users encounter when using computers and found that the most frequent causes of frustration are error messages, dropped network connections, long download times, and hard-to-find features, while the time lost due to the frustrating experiences ranged from 30.5% of time spent on the computer to 45.9% of time spent on the computer. The second study looked at psychological and social perspectives on frustration in an attempt to clarify the relationships among variables such as personality types, cultural factors, goal attainment, workplace anger, and computer anxiety and to develop a technology frustration model. The third study, still in progress, deals with subjects that use computers as a part of their work-related responsibilities. Specific goals of the study include: to determine the most frequent causes of frustration, to measure the time lost due to frustrating incidents, and to describe the impact that the frustrating incidents have on the users and their interactions with co-workers. The study will allow a better understanding of how user frustration impacts on employees and companies.

Our study

Designing the study

The study was triggered by Carol Hert's paper "Developing and Evaluating Scenarios for Use in Designing the National Statistical Knowledge Network" [24]. As part of the NSF Digital Government Initiative, Hert's project deals with a scenario-based approach to statistical information. In a first phase, fifteen scenarios were developed via a participatory design approach that involved agency partners. Team members of the project searched on the web the answers to these scenarios, in an attempt "to understand the specific integration challenges represented by a given scenario, the information that was available to address the scenario, provide further information on how the scenario might be better formulated, and gain insights on how existing sites were addressing integration challenges […] and how they were presenting data". As a result of this search, the team developed a database containing keywords used in the search process, URL's of the pages containing relevant data for the scenarios, types of statistics found, etc.

Having as starting point these 15 scenarios, we chose 3 of them, which we felt represented 3 hierarchical levels of need of information:

- Construct an understanding - the question as well as the answer is very elaborate and the source of information to be searched is not clear from the context
  Scenario: "I'm a social activist in the Raleigh-Durham, North Carolina area and have become increasingly concerned about urban sprawl and the loss of rural areas for both farming and recreation. I need statistics to support my claim that significant differences occur when urban development occurs in rural and/or farming areas."
- Search for specific data – the user only needs to locate information
  Scenario: "I would like to open a grocery store specializing in organic products in the greater Seattle metropolitan area. What are the trends in production and consumption of organic food products? Would the Seattle area be a good place to locate?"
- Comparative search - the user has to look for information regarding an interaction between two phenomenon
  Scenario:" I'm contemplating a move from Seattle to Bozeman, MT. How do they compare?"

Running the study

Previous to the study, we conducted a pilot test, to verify the correctness and usefulness of the procedure we intended to follow. The pilot study helped refine the observation methodology and provided a list of most common and frequent types of frustrations the subjects might encounter during the study.
We had 15 subjects, 6 females and 9 males, with different backgrounds: Computer Science, Library and Information Sciences, Economics, French, Sociology, Electrical Engineering, MBA and Medical Studies.
The study was conducted with two sets of subjects. The first 8 subjects were asked to navigate the FedStats web site using only the Topic A-Z links page. The other 7 subjects used as a starting point the FedStats homepage, and were allowed to use all the features of the web site.
All the subjects were given the three scenarios mentioned above. They were asked to find the answers to the scenarios using the FedStats web site and they were given 10 minutes for each scenario. A think aloud protocol was used, in order to observe and register the subjects' actions and comments.
After each scenario, the subjects were asked to fill out a short questionnaire intended to reflect their opinions about the scenarios and the results they got, about the web site ease of use and usefulness and about the level and type of frustration they experienced during the study.

Analyzing the results

Out of the total 45 times the queries were asked, 31 times the subjects did not find an answer. Only 7 times the subjects found the correct answer, while the other 7 times a partial answer was found.
For the first set of 8 subjects, in 16 cases no answer was found and in 3 cases the correct answer was found. For the second set of 7 subjects, in 15 cases no answer was found, and in 4 cases the correct answer was found.
These results indicate that there was no significant improvement in the subjects' performance when the entire web site, instead of just the Topics A-Z page, was used.
All subjects spent on average 7.13 minutes to look for information for each of the scenarios.

When asked about how satisfied they were with the way the query was stated, the subjects reported being somewhat satisfied or very satisfied in 28 (62%) of the cases (see Fig.1).

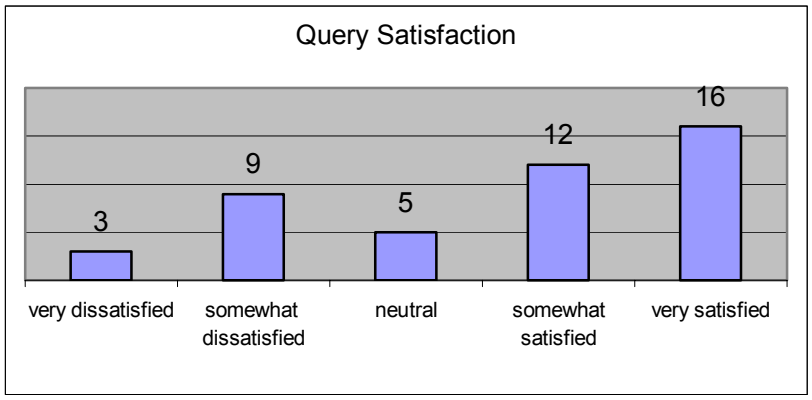

Fig.1 Query satisfaction results for all 45 instances of the study.

When asked about how satisfied they were with the way the answer found, the subjects reported being somewhat dissatisfied or very dissatisfied in 27 (60%) of the cases (see Fig.2).

Fig.2 Result satisfaction results for all 45 instances of the study.

When asked about how useful the FedStats web site was for them, the subjects reported that the web site was not useful at all or not very useful in 23 (51%) of the cases (see Fig.3).

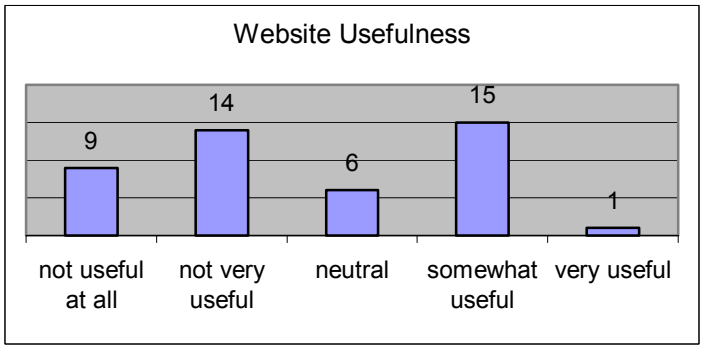

Fig.3 Web site usefulness results for all 45 instances of the study.

The rest of the questionnaire tried to assess the level of frustration the subjects experienced while trying to accomplish each task. They were first asked about the general level of frustration after completing the task, and then they were asked if they encountered specific frustrations and at what level. We provided a list of 7 frustrations that seemed to be most common during the pilot test, and we also asked the subjects to list any other type of frustration that might have occurred during the study.

The subjects reported general levels of frustration of 5 or above in 32 (71%) cases (see Fig.4).

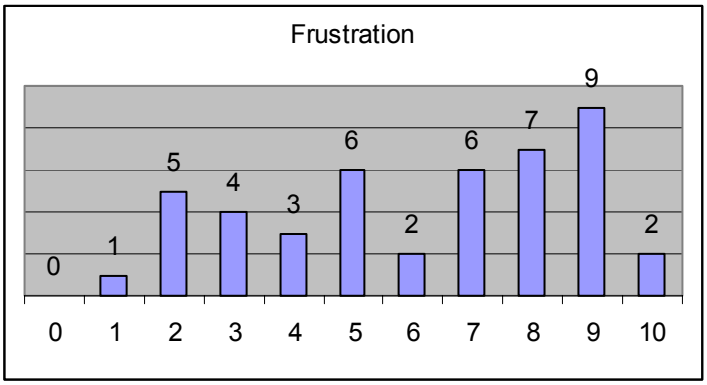

Fig.4 Frustration level for all 45 instances of the study.

64% of the subjects had no problem understanding the queries, as they were stated (see fig. 5)

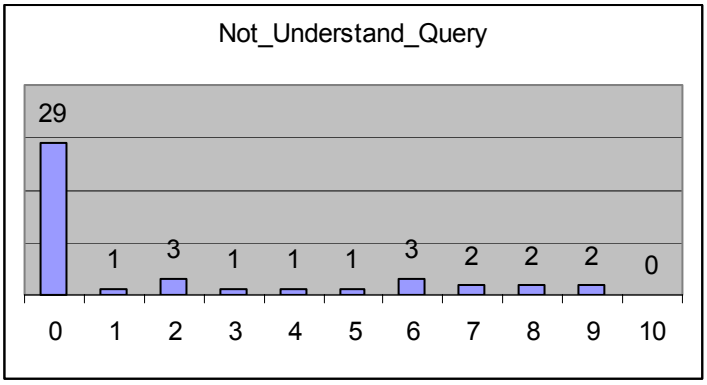

Fig. 5 Results from the statement: " I could not understand the query"

53% of the subjects reported no frustration in understanding the results. Their comments indicate that this is because most of them did not find any results that were relavant to the query. (see fig.6)

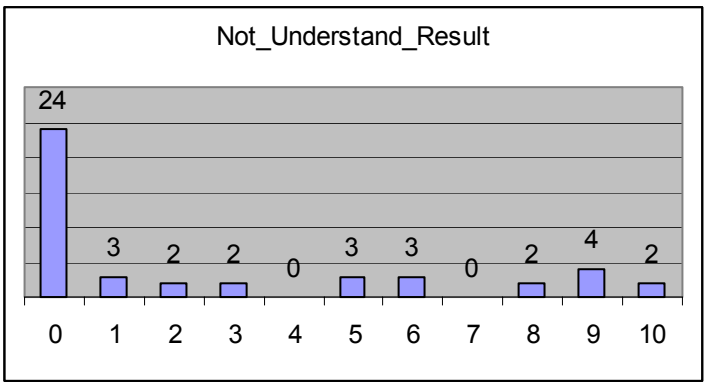

Fig. 6 Results from the statement: "I could not understand the results I got"

55% of the subjects reported a level of frustration of 5 or above related to the usefulness of the results they got. (see fig.7)

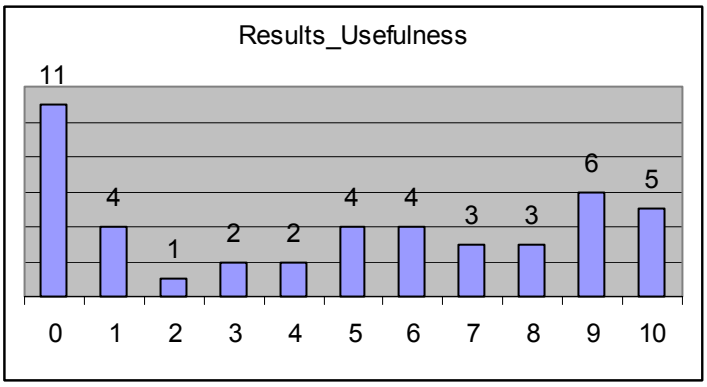

Fig. 7 Results from the statement: "The results were not very useful"

46% of the subjects reported a level of frustration of 5 or above related to the granularity, in terms of geographical level and time period, of the results they got (see fig. 8)

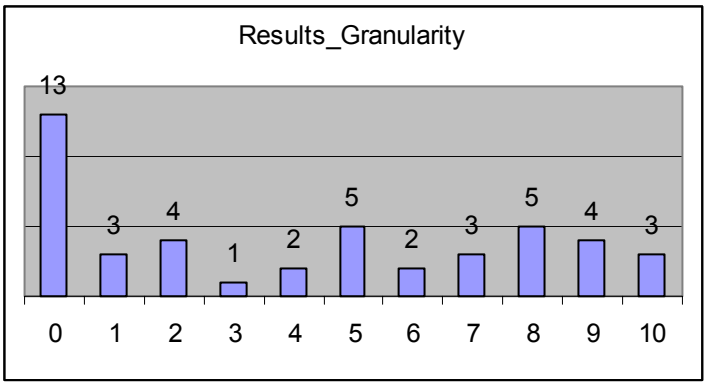

Fig. 8 Results from the statement: "The granularity of the results was not satisfactory"

Most of the subjects who came very close to the answer were not able to get a better granularity of the data because the refinement process was either hard (the ways to perform the refinement were not obvious) or impossible to do (there was no way to perform the refinement).

57% of the subjects reported a level of frustration of 5 or above related to how easy to use the web site was (see fig. 9)

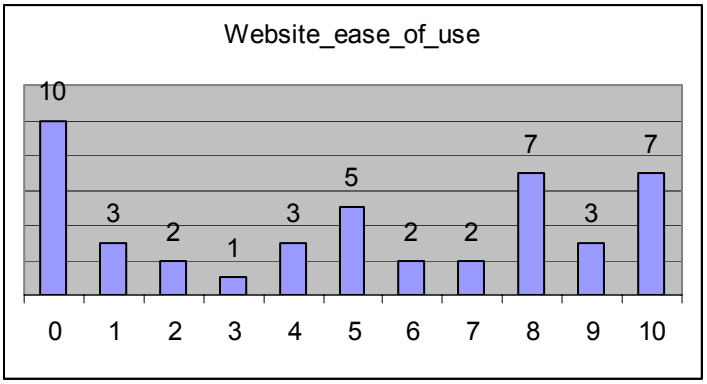

Fig. 9 Results from the statement: "The web site was hard to use"

The subjects complained that the language used on the web site was often not an appropriate one. Terms like "cities" were disguised under "metropolitan areas", making the search process hard for a user without previous knowledge of the statisticians terminology. Also, the subjects expected to be able to use tools that would simplify their task. For example, when wanting to compare attributes of 2 different regions, subjects could not perform a comparative search or a quality of life analysis. Even a simple search proved sometimes to be very ineffective and misleading, since logical operators were not always implemented properly in the search function.

48% of the subjects reported a level of frustration of 5 or above related to how confusing the searching process was while navigating the web site (see fig. 10)

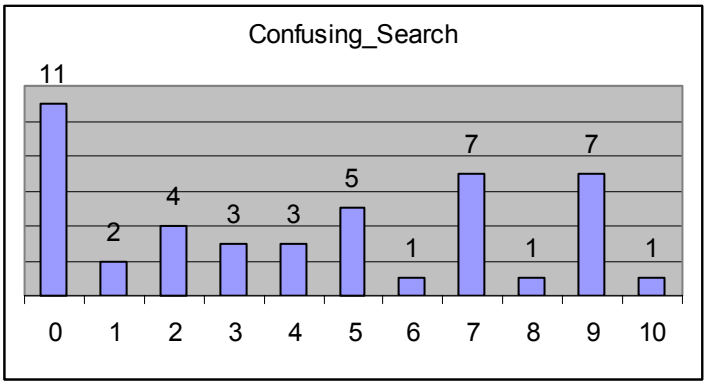

Fig. 10 Results from the statement: "The search process was confusing"

The subjects had a hard time navigating through the web site. The categories through which the data was made available did not prove too useful, and the subjects kept coming back and forth on the same paths, not knowing which one was actually relevant to their questions.

57% of the subjects reported a level of frustration of 5 or above related to how much time they spent searching for the correct answer (see fig. 11)

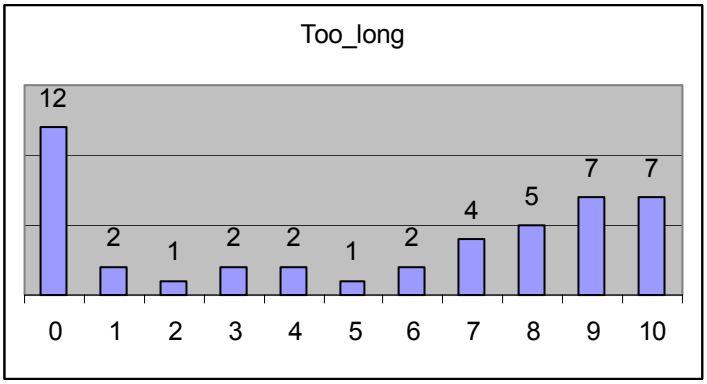

Fig. 11 Results from the statement: "I spent too much time"

The way the web site was structured and presented proved to be a challenge to the subjects. All the subjects were first time users of the FedStats web site and most of them were novice users, as far as statistical web sites were concerned. The time they spent on the web site was longer than expected, because the interface was not easy to learn and the search sometimes presumed previous knowledge of statistics.

The following types of frustrations were also reported by the subjects:

- Need more related links of the type "see also"
- Could not find links to any keyword in the query
- Need geographic granularity by cities
- Need search keyword option on Topics A-Z page
- Obvious keywords missing
- No "cost of living calculator"
- No way to set up comparative statistics
- Use easier to understand language
- The search engine does not always correctly implement logical operators
- Topics were confusing

The subjects were asked to freely talk about the experience with the web site. Here are some of their comments:

- Having a background or familiarity with this kind of research would probably help greatly
- FedStats was close to useless
- There is too much data
- Can I go to Google?

Examples of subjects' experiences

The following is an example of a subject's experience with the web site in trying to find the answer to the second scenario and failing.
The subject starts on the "Topic links A-Z" page, and looks for the following keywords suggested by the scenario: organic, food, Seattle, regions. Neither of them is found. The closest match was found in the link "Food consumption patterns", which the subject follows. The home page of the US Department of Agriculture Food Survey Research Group is displayed, and the subject chooses to follow the link "Foods Commonly Eaten in the United States, 1994-96" and performs a "Find in page" search for the keyword "organic". Since no results are obtained, the subject abandons this path and goes back to the "Topic links A-Z" page, and follows the link "Food stamps". The home page of the US Department of Agriculture Food and Nutrition Service is displayed, but the subject cannot find anything relevant on the home page, so this path is abandoned too. Finally, the subject goes back to the "Food consumption patterns" link, follows it once more, this time looking at the site map. The site map did not prove to be very helpful, and after consulting it for a while the subject chooses to follow the link " Food Consumption Surveys in the US Department of Agriculture". The document did not contain any statistical information related to organic food, so the subject gives up the path and the search process entirely.
To be noted that the subject did not find a way to restrict the search to the Seattle region. Also, since no clear link to organic foods and markets was found, the subject got lost in the huge amount of web sites and data that might have been relevant.

The following is an example of a subject's experience with the web site in trying to find the answer to the third scenario and succeeding.
The subject starts on the "Topic links A-Z" page, and looks for the following keyword (even though not suggested by the scenario): metropolitan areas. The link is found, and it leads to the US Census Bureau page about Metropolitan Areas. The subject finds a link to the "American FactFinder" web site and follows it. A number of links are available, providing information about: General Characteristics: Population and Housing; Economic Characteristics: Employment, Income, Poverty; Social Characteristics: Education, Ancestry, Language; Housing Characteristics: Value, Rent, Owner Costs. All data can be selected for a specific city or town. The subject stops the searching process, since all the desired data were found.
To be noted that the subject knew that statistical agencies use the term "metropolitan areas" instead of "cities". Most of the subjects did not find an answer to this scenario because they could not find a link related to "cities". Also, the link to "American FactFinder" seemed familiar, and the subject went directly to it. Most of the subjects who followed the same path clicked on all the links on the US Census Bureau page, except on the "American FactFinder" one.

Conclusions and Recommendations

Accessibility and universal usability should be two of the primary concerns of the governmental agencies when it comes to citizen services provided through the agencies' web sites. Dissemination of statistical information should be governed by the following design principles:

1. Universal access - The interface should accommodate the diversity of users: not only expert users, but also first-time and one-time users should be able to easily access and find the desired information. In the case of FedStats, we found that most of the subjects were confused by the design of the web site, and even after the second task, they did not "learn" the interface.
2. Easy Navigation - The information available should be presented in a structured way. All subjects indicated that a home page containing categories by topics, categories by agencies and a search function would best serve their needs. An alphabetical list of topics slows down the search process, especially when it cannot ever be complete from the user's point of view. In the case of FedStats, subjects indicated frustration when searching the A-Z topics without finding keywords they were expecting to find. On the other hand, the most used feature besides the search function was "Search by agency".
3. Common language - The terminology used to present the information available should be easy to understand. Except for the expert users, all others most likely do not have enough knowledge to look for the "scientific" term, and instead use everyday language words. Also, the agencies should not expect the users to know the structure, the exact role of each agency, or the interactions between agencies. All these should be transparent in the search process. In the case of FedStats, many users complained about the keywords used on the web site (e.g. the common keyword "cities" is hidden under the not so common phrase "metropolitan areas"). On the other side, the subjects found very helpful the "key statistics" link and the short description for each agency.
4. Comparative search - The web site should allow comparative search. Since this is a very common task for users searching statistical information, the web site should provide ways to comparatively view and analyze data. In the case of FedStats, the third scenario proved to be more difficult and more time consuming than it should have been because the subjects had no way to perform a comparative search.
5. Advanced search - The search feature should have full functionality. It should support a comprehensive search through the huge amount of data available, support logical operators and provide relevant output. In the case of FedStats, although the search box was the most commonly used method to find the answer to the scenarios, in most cases it provided useless output, and sometimes misled the subject by not correctly implementing the use of logical operators.
6. Data Tools - Common-use ways of viewing and analyzing statistical data should be available on the web site. Statisticians and researchers need to be able to quickly perform analysis of data by certain well-established criteria. In the case of FedStats, most of the users were unpleasantly surprised to notice the lack of a "cost of living calculator" feature.

7. Data Granularity - Allow users to choose the granularity of the information searched in terms of geography and time. In the case of FedStats, subjects were often not able to find the information at the city level, being offered data only at the state or county level. Additionally, subjects expressed the desire to be able to choose the time interval for which they want to search for data.

To summarize, this study was intended to discover the advantages and the shortcomings of a web site that provides access to statistical information. We chose the FedStats web site, since it is the biggest portal to governmental statistics on the Internet. While our usability study had a reduced scope, it provided sufficient insight and experience with this type of analysis to lead to a future study of larger coverage.

Acknowledgements

We appreciate the support from the National Science Foundation grant for Information Technology Research (#0086143) Understanding the Social Impact of the Internet: A Multifaceted Multidisciplinary Approach.

References

1. Bosley, J., Straub K. - Data exploration interfaces: Meaningful web database mining by non-statisticians, IBM Make IT Easy 2002
2. Ambite, J. L. et al - Simplifying Data Access: The Energy Data Collection (EDC) Project, IEEE Computer, February 2001, pages 47-54
3. Donohue, K. – Designing Interfaces to Statistical Databases, CHI Workshop Interacting with Statistics, 1999
4. Johansen, P. - Exploring Regional Statistics, CHI Workshop Interacting with Statistics, 1999
5. United Nations Statistical Commission and Economic Commission For Europe – Guidelines for Statistical Metadata on the Internet, Conference of European Statisticians Statistical Standards and Studies No. 52, Geneva, 2000
6. United Nations Statistical Commission and Economic Commission For Europe – Best Practices in Designing Websites for Dissemination of Statistics, Conference of European Statisticians Methodological Material, Geneva 2001
7. Hert, C. A., Marchionini, G. - Seeking Statistical Information in Federal Websites: Users, Tasks, Strategies, and Design Recommendations: Final Report to the Bureau of Labor Statistics. (July 18, 1997).
8. Marchionini, G. - Advanced Interface Designs for the BLS Website: Final Report to the Bureau of Labor Statistics. (June, 1998).
9. Hert, C. A - FedStats Users and Their Tasks: Providing Support and Learning Tools: Final Report to the United States Bureau of Labor Statistics. (August, 1998).
10. Hert, C. A - Federal Statistical Website Users and Their Tasks: Investigations of Avenues to Facilitate Access: Investigations of Avenues to Facilitate Access: Final Report to the United States Bureau of Labor Statistics. (1999).

11. Marchionini, G. - An Alternative Site Map Tool for the FedStats Website: Final Report to the Bureau of Labor Statistics. (June 30, 1999).
12. Haas, S. - Knowledge Representation, Concepts, and Terminology: Toward a Metadata Registry for the Bureau of Labor Statistics. Final Report to the United States Bureau of Labor Statistics. (July, 1999).
13. Haas, S. - A Terminology Crosswalk for LABSTAT: Mapping General Language Words and Phrases to BLS Terms: Final Report to the United States Bureau of Labor Statistics. (September 29, 2000).
14. Haas, S. - From Words to Concepts to Queries: Helping users Find Series and Variables to Satisfy Their Information Needs: Final Report to the Bureau of Labor Statistics. (November 12, 2001).
15. Hert, C., Marchionini, G., Liddy, E., and Shneiderman, B. (2000). Interacting with tabular data through the World Wide Web. White paper presented to the FCSM Statistical Policy Seminar, November 2000.
16. Hert, C., Marchionini, G., Liddy, E. and Shneiderman, B. (2001). Integrating electronic systems for disseminating statistics. Federal Committee on Statistical Methodology Statistical Policy Working Paper 32: Seminar on Integrating Federal Statistical Information and Processes. (April 2001). Washington, DC: Office of Management and Budget. 219-226.
17. Marchionini, G., Hert, C., Liddy, E., and Shneiderman, B. Extending Understanding of Federal Statistics in Tables. *ACM Conference on Universal Usability*. (Washington, DC). 132-138.
18. Liddy, E.D. and Liddy, J.H. (2001). An NLP approach for improving access to statistical information for the masses. Paper presented to the FCSM 2001 Research Conference, November 2001 (Abstract)
19. Marchionini, G., Hert, C., Shneiderman, B. and Liddy, E. (2001). E-tables: Non-specialist use and understanding of statistical data. *Proceedings of dg.o2001: National Conference for Digital Government.* (Los Angeles, May 21-23, 2001). 114-119.
20. The GovStat Project - Integration of Data and Interfaces to Enhance Human Understanding of Government Statistics: Toward the National Statistical Knowledge Network, 2002, http://ils.unc.edu/govstat/
21. Ceaparu, I., Lazar, J., Bessiere, K., Robinson, J. and Shneiderman, B. - Determining Causes and Severity of End-User Frustration (May 2002)
22. Bessiere, K., Ceaparu, I., Lazar, J., Robinson, J. and Shneiderman, B. - Social and Psychological Influences on Computer User Frustration (July 2002)
23. [Frustration paper no. 3]
24. Hert, C. - Developing and Evaluating Scenarios for Use in Designing the National Statistical Knowledge Network, 2002