Co-evolution of user and organizational interfaces: A longitudinal case study of WWW dissemination of national statistics

Gary Marchionini
University of North Carolina at Chapel Hill

Abstract

The data systems, policies and procedures, corporate culture, and public face of an agency or institution make up its organizational interface. This case study describes how user interfaces for the Bureau of Labor Statistics website evolved over a five year period along with the larger organizational interface and how this co-evolution has influenced the institution itself. Interviews with BLS staff and transaction log analysis are the foci in this analysis that also included user information-seeking studies and user interface prototyping and testing. The results are organized into a model of organizational interface change and related to the information life cycle.

Introduction

That information technology is strongly influencing all aspects of life today is a common refrain in both scholarly and popular venues. These changes extend to government service in many ways and may be organized over four general and non-orthogonal dimensions that change at different rates and interact to create higher order effects that in turn influence the main dimensions. The general dimensions are information technology, data, people, and organizations. Information technology includes hardware, software, and network facets. It is axiomatic that the technology itself is changing rapidly; faster, more powerful computers and higher-bandwidth connectivity regularly appear at the same price points as existing hardware. Although less frenetically-paced, new and improved software makes Internet access and usage easier and more effective; tools such as web browsers and associated interface styles evolve as the hardware infrastructures improve. Likewise, the bandwidth and penetration of Internet access continues to increase rapidly and new alternatives for wired and wireless delivery emerge. The data dimension likewise sees increases in volume as well as new forms and genres. Each day there is more information available to the public from more points of view. In addition to the entire gamut of media forms, new forms such as simulations and code libraries are available, along with a variety of aggregations, indexes, and metadata associated with these information resources. The people dimension is growing rapidly as larger and more diverse portions of the population use computers in daily life and have Internet access in their homes. Not only are more people with more diverse needs and experience using information technologies and digital data but the installed base of experience and expectations ranges across a much wider spectrum. Finally, the organizational dimension demonstrates similar growth in quantitative and qualitative parameters as organizations replicate existing services in the Internet environment and create new electronic products and services. Whether they are public or private, local or global, organizations today have adopted the Internet as a tool to achieve their missions and in many cases expand their mission and services. All these trends are well-documented, and seem likely to continue for the immediate years ahead. They especially apply to knowledge intensive enterprises such as government service at all levels where information is the primary resource. What is most interesting are the interactions among these dimensions, and by extension, the effects these interactions have on subsequent interactions. This paper examines these interactions and effects in a large federal statistical agency and presents a model of how information technology, data, people and organizations co-evolve.

The examination is based on interviews, transaction log analysis, and user interface designs and usability studies over a five-year period. This paper focuses on the interviews and transaction log analyses and the resulting co-evolutionary model. It incorporates aspects of well-known

phenomena such as human information-seeking behavior, human-computer interaction, the information life cycle, and technology adoption and organizational change. It also incorporates less well-developed phenomena such as statistical and technical literacies, interaction styles, multifaceted analysis methods, and organizational interfaces. The focal points for this co-evolution are the various intersections—interfaces--among the dimensions. In particular, the emphasis is on user interfaces and more general organizational interfaces manifested by the statistical agency.

User interfaces relate people to specific computer systems and represent an entire field of study in their own right. In general, user interfaces integrate principles from computer science, psychology, and information engineering (e.g., Norman, 1998; Shneiderman, 1998). User interfaces for information seeking and use are driven by these fields as well as by information retrieval research (e.g., Hearst, 1999; Marchionini & Komlodi, 1999). Much of the empirical evidence reported here was motivated by the goal of improving user interfaces for statistical websites.

By *organizational interface*, I mean the data systems, policies and procedures, corporate culture, and public face of an agency or institution. This is a broad usage of the term encompassing the intersection of people, data, tools, and policies specific to an organization. The public face component of the organizational interface is typically represented by WWW page designs and specific user interfaces and by the actions and demeanors of organizational staff who interact with the public. Malone (1987) used the term organizational interface to differentiate the user interface relating an individual to a computer from the "parts of a computer system that connect human users to each other and to the capabilities provided by computers. P. 294" Thus, the organizational interface is the basis for people to collaborate and for organizations to coordinate activities. Barreau (1997) used it to describe "the ways in which organizations influence the use of information and information systems. P. 12" Thus, the organization interface involves the policies and processes by which organizations manage and use information to achieve their missions. In today's electronic environment, organizations work hard to project mission and style through 'branding' that aims to incorporate the organizational and customer viewpoints (e.g., Bickerton, 2000). Thus, an organization's projected brand is also one important element of the organizational interface.

For the work reported here, crucial elements of the organizational interface are the user interfaces provided though websites. The user interface is the obvious entry point and 'image' of an organization from the user's perspective—the purposive brand projected by an organization. However, it is also the nexus of the hidden policies and perspectives within the organization as well as the data, and technology status of the organization. Interfaces project purposeful and latent 'faces' that determine user perceptions. Although clever design may temporarily project a wished for impression to users, over time, the underlying nature of an organization and the image projected through its interfaces must converge. Just as "an actor's mask eventually becomes his face", an organization's user interface eventually influences the overall organizational interface and in turn, the organization itself. This is particularly important for government or service-oriented organizations as WWW interfaces lead to more widespread usage. As the user base expands and becomes more diverse, the services, missions, and interfaces central to the organization will necessarily adapt if the organization is to sustain itself. This hypothesis is demonstrated in the case of the Bureau of Labor Statistics presented here.

Context of the Case

From the Fall of 1996 through 2001, a team of information specialists has been working with the Bureau of Labor Statistics (BLS) to improve outreach to constituent groups by means of the World Wide Web (WWW). One general goal of this work was to gain a better understanding of how non-specialists think about, access, and use statistical data. A second general goal was to understand and document how federal statistical agencies, the Bureau of Labor Statistics in particular, can adopt and adapt technologies to better serve the needs of diverse constituencies. Much of the work focused mainly on a specific goal to design and test user interface tools that help citizens understand what federal statistics are available, access the statistics that are most pertinent to their needs, and use these data to answer questions and make decisions. The user interface design vision was guided by the belief that people benefit from multiple views of information resources and dynamic mechanisms for manipulating these views (e.g., Greene et al., 2000; Marchionini, et al., 2000; Shneiderman, 1998). Furthermore, such designs must be informed by understanding people's needs and behaviors, and by rich organization and indexing in the underlying information. The results of these efforts are reflected in the changes the BLS website made over time and in the current organization and information-abundant design.

The BLS is an agency within the U.S. Department of Labor with the mission to be "the principal fact-finding agency for the Federal Government in the broad field of labor economics and statistics" (www.bls.gov). BLS conducts many periodic surveys such as the Current Population Survey that collects data from 50,000 households each month. The data collected in various surveys are used to prepare important economic indexes such as the Consumer Price Index (CPI) and Producer Price Index (PPI) as well as a multitude of reports on employment and other economic conditions. As a federal agency, BLS aims to provide broad access to data and was an early adopter of Gopher and ftp tools to disseminate data and by 1996 was providing WWW access as well. The current BLS website has evolved over the years and made use of the procedures and results of the research emanating from the team of information science scientists[1].

In our preliminary work (1996-97), we conducted investigations of user needs and tasks and completed transaction log analyses of user behavior at the BLS website, resulting in a user task-type taxonomy (See Seeking Statistical Information in Federal Websites: Users, Tasks, Strategies, and Design Recommendations http://ils.unc.edu/~march/blsreport/mainbls.html). In the second year (1997-98), additional user needs analyses were done, transaction log analyses were replicated, an interface prototype that aimed to provide alternative entry points to the BLS website for different user needs and types was built and tested, and recommendations for short and long term design strategies were made (Hert & Marchionini, 1998); (See Advanced Interface Designs for the BLS Website http://ils.unc.edu/~march/blsreport98/final_report.html). In the third year (1998-99), the interface work extended these efforts to the Fedstats website that includes the bulk of statistics collected and disseminated by 70 U.S. government agencies. Fedstats is a portal service (also called a locator service or a metasite) meant to serve as a gateway to all federal government statistics (www.fedstats.gov). A prototype tool (Relation Browser) was built that aims to give people an overview of the range of federal statistics and an alternative entry to them through "look ahead" interface mechanisms. This prototype was tested

[1] The original team was composed of Carol Hert and Gary Marchionini and was expanded to include Stephanie Haas in 1998. After the first year, each team member focused on different aspects of the research agenda and prepared individual reports and papers. Hert focused on continuing to examine user needs and behaviors and examined metadata issues. Haas worked on vocabulary analysis. All annual reports are available on the respective researcher websites (http://istweb.syr.edu/~hert/ and http://www.ils.unc.edu/~stephani/ and http://ils.unc.edu/~march/)

with users and recommendations for revisions made (Marchionini et al., in press) (See An Alternative Site Map Tool for the Fedstats Website http://ils.unc.edu/~march/blsreport99/final.pdf). In the fourth year (1999-2000), the Relation Browser was revised based upon the previous year's usability tests, two types of assessment were made, and a final revision was prepared. (See From Overviews to Previews to Answers: Integrated Interfaces for Federal Statistics http://ils.unc.edu/~march/bls_final_report_99-00.pdf). In the final year (2000-2001), a replication of the transaction log analysis and interviews with BLS staff was undertaken to examine how the BLS data dissemination effort had evolved and make recommendations for future interfaces. This reflective analysis is the main basis for this paper. The paper first presents results from the interviews, then discussed the uses and limitations of transaction logs and presents results over the five year period, and finally discusses these results in the context of organizational change and the co-evolution of the primary dimensions that interact to define these changes.

Interviews

In the Fall 2000, nine formal interviews were conducted with BLS staff. Four of these interviews were with people who were interviewed in 1996-7. Of the nine interviews, six were conducted in person and three were conducted over the telephone. In all cases, handwritten notes were made, and in the face-to-face interview cases where the interviewee agreed, an audio recording was also made. In addition to these formal interviews, several informal discussions were held with senior managers at BLS.

As in the 1996-7 interviews, a structured protocol was used to guide the discussion. Questions first aimed to establish a context for the interviewee's roles at BLS, then focused on the user community. These questions aimed to establish what types of people used BLS services (e.g., occupations), what types of information they sought, and the volume of questions and what forms they took (e.g., phone, email). The final set of questions probed the perceived changes in service from the public's point of view and then the changes that the interviewee saw in their own jobs at BLS and in the overall operation of BLS as an institution. Interviewees were also encouraged to make suggestions and comments about improvements or issues related to WWW technology at BLS.

The results from the interviews are organized to first provide a summary of the user population and then summarize changes WWW technology has brought for the public, for BLS employees, and for the institutions. Of the interviewees, one was mainly responsible for backend data system operations and had little direct interaction with the public, four were analysts in different BLS program areas who interacted with the public on questions directly related to a survey or program, two were directly responsible for providing data to the public and responding to requests, and two were managers who were had some interactions with the public but mainly through their staff.

BLS website users

Although some of the interviewees saw little difference in the users in 2000 than in 1996, there were strong statements about broader and more diverse users of BLS services. Long-time user groups that were specifically mentioned by interviewees include journalists, academics, state government staff, students, and legal aides looking for contract escalation data. New audiences were exemplified in several ways. One interviewee noted requests for help on applying for unemployment benefits, information that is not available at BLS but rather from its parent Department of Labor and state or local agencies. Another made comments in a similar vein, noting that there were questions that show that people have no idea about what BLS provides—comparing these requests to requests arriving by letter where people knew a great deal about what

data BLS collects. One noted an increase in K-12 student requests. Another noted that people tend to request information about local information, whether it is available or not. An interviewee with help desk responsibilities noted a wide range of users with a range of requests. Another user noted that some email requests are arriving in Spanish. It seems clear that BLS is serving an increasingly diverse set of users.

The questions that people ask also reflect this increasing diversity. On the one hand, regular users with specific requests continue to use BLS services, for example, the legal aids, accountants, state/local representatives come back each month for one or a few specific value(s) and represent a fairly stable class of users. Another traditional group of users is the set of researchers from academic institutions, Congressional offices, think tanks, and the media. These users want in-depth data and are relatively sophisticated in their understandings of what BLS collects. The largest and newest group of users are first-time or casual users with questions that are often geographically localized (e.g., information about salaries for a particular occupation in a specific city) or personalized (how to get my birth certificate, how to file for unemployment).

The 2000 BLS Customer Service Guide (BLS, 2000a) notes BLS receives over 30,000 requests per month. Requests come to individual departments or people as well as to central information centers that route requests. One thing that all interviewees agreed about was the increase in email requests. Most noted that phone requests have decreased, although not as much as email requests have increased. One interviewee estimated 900-950 emails per month with the volume slowing down in the summer. Another noted that the number of letters has decreased and phone calls have gone from 5000 to 2000 per month over a seven year period, while emails have increased to several hundred per month. Another noted that the predominance of phone calls (about 500 per month) is dropping slightly while emails range from 50 per month in slow months to 70-80 per month other times. These wide differences in numbers reflect the different jobs and program areas of the interviewees, but in all cases, show a rising volume of requests and a broadening of formats with email representing the biggest growth.

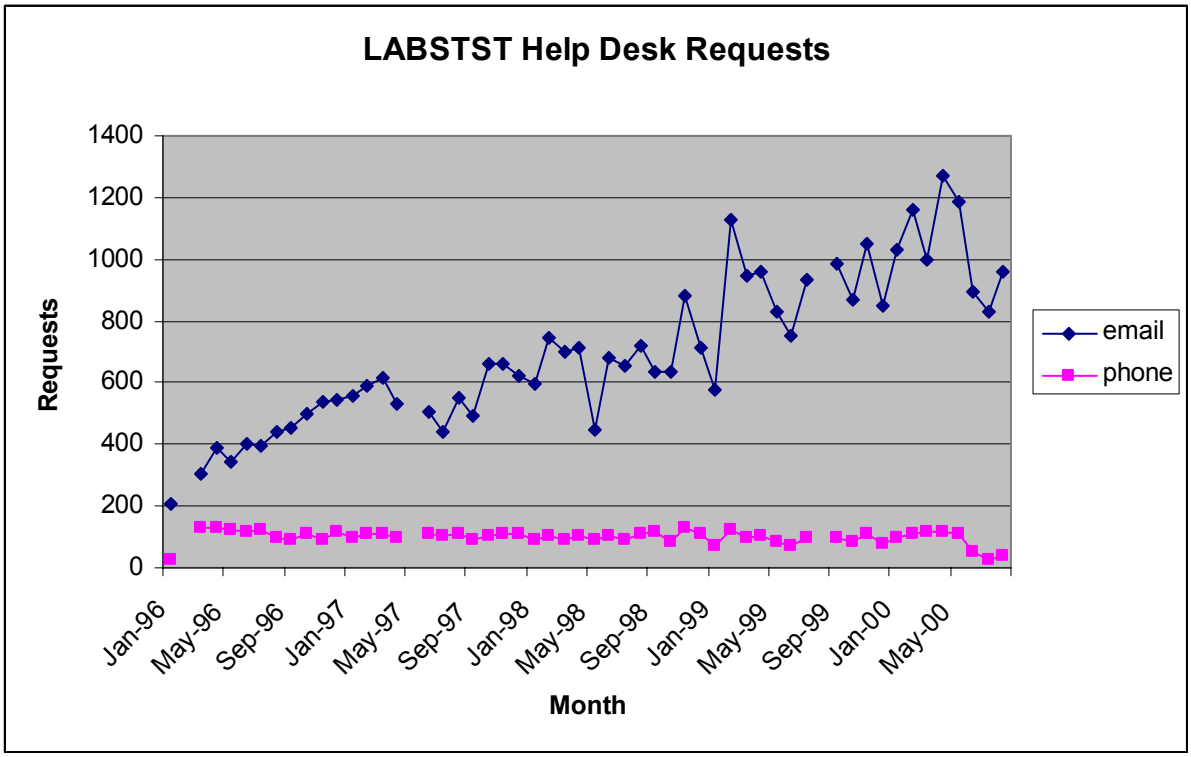
The LABSTAT unit is responsible for the public website and is thus a focal point for contacts. Figure 1 displays data provided by the LABSTAT helpdesk for a 56-month period. We would expect that many of the requests to its help desk would be related to the website and therefore be made by people with computer access. Thus, the email request volume over phone volume is not surprising, however the upward trend in email volume with relatively constant phone volume is of interest and reflects the verbal estimates made by interviewees across BLS departments

It is useful to consider interviewee comments about how user requests, especially email are managed. One noted that the department policy is 24-hour turnaround for all requests with a goal of one-hour response. This is an extraordinary level of service. Several noted that email responses from BLS are treated as official documents and thus require careful construction and in two cases, review by a manager. This is a significant level of effort beyond what might go into a phone response, assuming the time to find the answer is the same.

Changes Due to the WWW Interfaces: Users. In responding to prompts about how the BLS website has affected the public, most interviewees noted three improvements that represent a technical advantage theme: faster access to wider arrays of data at all times of the day. These are commonly trumpeted advantages of the WWW in all venues and they apply at BLS perhaps even more than e-commerce and other services. The overall effect is to be more 'customer' centered—provide what the customer wants, whenever they want it, with minimal delay. These are powerful competitive advantages for business and it is worth considering their advantages to non-profit institutions like government agencies. These obvious advantages underlie secondary

changes that in turn cascade back on customer services. In this regard, a second theme related to user expectations emerged. Three of the interviewees specifically said that user expectations have increased. In addition to expecting instant access, they cited different examples where WWW access raised demands: People want historical data that parallels new data, even when this data was not collected until recently. They want instant answers tailored to their questions rather than simply being directed to the website. They want 'all' the data (including archived data). They want localized data. These examples illustrate on the one hand a less knowledgeable user population in terms of understanding what BLS does, but on the other hand, users who are becoming more "data aware" (to use one of the interviewee's terms) and are exploring the limits of their new-found access. Interviewees pondered the new demands these rising expectations will bring to BLS. For example, because unrealistic expectations are unlikely to be met, will customer satisfaction decrease? Will higher expectations cause new demands for information that is not now collected?

Figure 1. LABSTAT Help desk Requests



A third theme that emerged in the interviews was the rising usability of the website. One interviewee noted that through the revised website people were offered entry points below the top, thus adding flexibility. Two interviewees noted improvements in vocabulary usage (more popular terminology, less arcane codes and agency jargon) on the website, although they both noted that there were significant improvements to be made in this regard. Another noted the importance of considering usability throughout both the data collection and the dissemination processes. Another interviewee noted that "it is easy to put things on the Internet, but not to do it correctly" in describing the need for user-friendly access as well as data security. Overall, these interviewees agreed that citizens are receiving better service (one noted that "data users love the WWW") but also raised questions about how these improvements can be maintained and extended, especially in light of rising expectations. This surely is a crucial issue for BLS and other agencies to consider in strategic planning. One strategy might be to aim to provide a basic

level of access to the broadest range of potential users, thus plateauing depth of service (e.g., not adding new types of data but making existing data more accessible); another is to continue to respond to the increasing expectations and needs of the new users (e.g., by adding new types of data and tools as these users expertise grows). Clearly, this in not an either or choice and decision makers must find ways to balance these competing goals within the constraints of limited resources.

Changes Due to the WWW Interfaces: BLS staff. The interviewees were forthright and thoughtful in expressing their feelings about how their own jobs have changed over the past half-decade. They spoke confidently in response to prompts because these were things that directly and continually affected their own lives. They made the following observations. The WWW has forced employees to upgrade their technical skills—examples ranged from web-specific skills like HTML to email and other generic applications. One suggested that many employees have gone from clerical careers to web careers. One noted that there was much more time spent using a computer and much less time spent using the telephone. One said that employees were becoming more world-oriented (rather than internally oriented) due to the WWW. One noted that improved communication (e.g., Intranet sharing, email, elists) makes work more efficient, but also noted that the backend database work has not changed. Another praised the specialized internal tools on the Intranet (e.g., Java applications) that helped one to find and use information internally. Another said that the creator of a new product must consider how the product will work on the web. One noted that morale in the organization was extremely high as a result of these changes. These observations are highly encouraging and may be considered to be a general effect of technology adoption within a knowledge-industry.

Other themes were also apparent. Multiple interviewees addressed growing personal accountability. One discussed this in terms of data accuracy and preparing data for the web—"now your work is available to the world with your email attached!" Several others noted that there were new levels of review for their work. Two mentioned email responses to the public as examples of work that is reviewed by others before sending. Another noted the process for reviewing papers and reports before posting to the website. This personal accountability theme has a strong institutional counterpart discussed below.

A related theme that arose in many interviews was security and the criticality of data release schedules. One manager noted devoting considerable time to release time management. One interviewee noted that there was more data to time manage. Another addressed the needs to manage who has updated what file at what time. A fourth interviewee noted that "four years ago, late was the problem, today it is too early." The interviewer noted the physical changes at BLS in this regard. Over the five years, there have always been security procedures for entering and exiting the building, however in the 2000 interviews, there were also locked doors beyond egress/exit points to entire suites of offices associated with data analysis for an upcoming data release. Clearly, security and data release management on the website have become high-priority issues for BLS and its employees.

In sum, it is clear that work at BLS has changed due to Internet technology. These changes are reflected in the actions people take as part of their jobs, what is valued at BLS in the organization, and the attitudes employees adopt. Overall, BLS is a knowledge-intensive agency and like counterparts in government and industry, employees have strongly adopted information technology. They have more immediate contact with the world beyond BLS, have new levels of accountability and are more aware of data security and the impact their work can have beyond the agency. Overall, these changes may add new pressures and higher expectations, but also tend to reinforce an esprit de corps that is a positive sign for government service.

Changes Due to the WWW Interfaces: BLS as an Institution. Changes wrought through technology to individual work propagate to changes in the institution just as changes in the institution affect individual work. In the case of Internet technology, BLS, like other institutions was more strongly influenced in the early days in a bottom up fashion than through top-down policies. Departmental or functional web pages within the agency appeared before there was an entity-wide website. One interviewee hinted at this with the comment: "There is a big difference in understanding between what is possible technically and what upper management understands—they know that technology can bite them, but not how or why." Clearly, the intense media attention given to premature releases of the CPI and effects on the stock market have driven home one of these effects and caused much of the attention to security and release management. In effect, the costs of early release have gone up dramatically due to the WWW as more eyes have immediate and easy access to the data. Some of the other challenges that lie ahead are hinted at in the themes discussed above. These changes and associated challenges fall into three categories: customer service, quality control, and information infrastructure and in all cases we can detect a maturation of theory and practice.

The Internet has caused government, academe, and industry alike to become more "customer" oriented. In the case of government, the substantial efforts toward digital government that began with government services through the Internet has blossomed into formal programs and initiatives (e.g., legislative edicts, professional conferences, scholarly publications, funded research groups). Citizens increasingly expect to not only access government information through the Internet but also conduct transactions e.g., (taxes, licenses, etc.). The changes implied by this customer orientation are myriad. Decisions about what data to provide and how to collect and provide more data were mentioned in the interviews at BLS. Consideration of more diverse user populations arose repeatedly in the interviews and in various meetings and publications. It is clear that BLS must continue to improve user interfaces to its public-access systems, walking a fine line between technical capabilities and the evolving installed base of systems and technical and statistical literacy among the population. The information architecture issues include mapping common and technical vocabularies, organizing the plethora of services and data, providing alternative interaction styles, and inventing new transactional services.

Another facet of customer service that is evolving quickly and came up repeatedly in interviews is how to manage the increasing volume of requests that come in via letter, fax, phone, and email. Email is especially crucial since it is the fastest growing medium of agency-citizen interaction. Over time, policies for email response in the different units have evolved, with fewer ad hoc procedures and more formalized procedures. Several interviewees reported on email routing schemes (alternating time schedules or functional delegations) and at least one reported a two-level review process before email responses were sent. Just as customer-service management (CSM) procedures for phone calls have evolved over the years, CSM procedures for email are under development and testing.

Quality control is an umbrella concept for a set of issues that have become increasingly evident over the five year period. Issues of data security and release timing have been addressed above and surely are the most high profile institutional changes due to the Internet. An equally important issue, however, deals with the confidence the pubic has in BLS data itself. This is perhaps a more crucial issue in the long term as it goes to the heart of BLS' mission. One interviewee noted that people trusted information on the BLS website more than information provided by BLS personnel on the phone. Whether this is actually the case or not is unclear, but surely there is a trust on the part of most citizens that information found on a government website is accurate. Two other interviewees talked about challenges to data quality brought about by

web-based dissemination. One noted that in the mainframe days when there were a small number of users, corrections could be made and sent to the users directly. Today, updates and corrections must be made to many different pages and there is no hope of alerting all those who have already accessed the early data. Another interviewee discussed the problems that corrections bring to programs that deal with state or local agencies and the need for more data checking. This interviewee went on to note that because data are released in multiple formats (multiple paper and multiple electronic), update and correction management are even more challenging. These concerns cause one to begin to wonder about whether there are increasing levels of noise/error in the public record, how corrections are propagated, and how the public's perceptions and understandings of data accuracy will evolve over time.

Quality control has emerged as a significant issue over the 1996-2000 period at BLS. There were a few comments about this in the early interviews (some discussion of email as public record that needed review, and concerns about data confidentiality in 1996 that did not emerge in the 2000 interviews). Discussions about review of email, web page tables and other products, and formal reports were common in the 2000 interviews. Surely, BLS (and DOL) as an institution will eventually evolve policies that address these issues.

Finally, there are changes at BLS is how the IT has become part of the information infrastructure. Although interviewees pointed out that the BLS Intranet is not the only internal communication and information transfer channel and the Internet is only one aspect of the dissemination process, it is clear that the these facilities have become institutionalized. One interviewee in discussing quality control, stressed the importance of quality control subgroups that test new backend systems. The visibility the website brings to BLS also drives changes such as devoting resources to building and testing more usable interfaces. The BLS website is not viewed as a static service for posting new data but as an evolving service with scheduled redesigns and testing. The Internet systems at BLS are no longer novel or experimental, but rather crucial elements of the data flow from collection to use. Issues to address range from practical issues such as how to manage paper and electronic services, to thorny system interoperation issues like how to integrate mainframe backend systems with internal and external Internet systems, to long-term issues like the implications of end-user access and behavior for creating new surveys and data services. These issues represent the interdependence of the user interface expressed via the BLS website and the larger organizational interface to which it provides entry.

BLS has come a long way from serving the needs of a few dozen companies, research institutions, and government agents who obtained tapes of data sets to a broader group with sophisticated computer skill and systems who used the Internet to transfer datasets via ftp, to today's web-based dissemination to anyone with a personal computer and Internet access. The interviews illustrate that there is a noticeable maturity and seriousness about the BLS website and other Internet-based technology. With this maturity come new sets of issues that leaders can anticipate by seeking the advice of front-line BLS staff as well as the public who use these expanding services.

Transaction Log Analysis

User behavior on the Internet can be assessed through surveys and questionnaires and there are a number of market research services that provide such services across the Internet as a whole or for specific sites and markets (e.g., Nielsen, ComScore, Netsizer). Another approach is to directly assess user behavior through analysis of electronic records specific to the institution itself. Electronic mail messages and transaction logs are two kinds of records used in these analyses. In the 1996 study, Carol Hert led the effort to content analyze email messages. Hert (1998) and

Haas (2001) have continued to examine email messages and query logs to investigate terminology issues. (see also BLS [2000b] internal report showing OCC and common requests such as inflation, unemployment, CPI and PPI as the most common requests people made). In the work reported here, the focus was on transaction logs as indicators of user behavior over the 1996 to 2000 period.

As the Internet has become a more mainstream vehicle for information flow in all aspects of life, business and government have sought better ways to manage and use the secondary data in the Internet infrastructure itself. In the case of the WWW, web server software typically records all incoming requests and system responses. These records are like all transaction systems and are necessary for system maintenance. In addition to basic system maintenance and reliability uses, managers also mine these 'transaction logs' to better understand user behaviors. It should be noted that transaction logs have been used to investigate user behavior for decades before the WWW (e.g., Campagnoni & Ehrlich, 1989; Marchionini, 1989; Penniman, 1975; Rice & Borgman, 1983; Tolle & Hah, 1985). Such understanding can, in turn, be used to provide better services and in commercial settings, gain competitive advantage. In fact, there are many market research groups (e.g., Nielsen, CommerceNet, comScore) that use sampling techniques and interviews as well as "click stream" monitoring (e.g., DoubleClick). Thus, transaction log analysis has become a mini industry as server software evolves to provide more options for logging and specialized analysis packages appear to process these logs. There are several systems involved in transaction log analysis, each having its own set of settable parameters that affect the overall analysis.

First, web server software allows system administrators to decide what to log. The basic logs include three types of information: access (e.g., date/time of request, IP address of requesting client, and page/program requested), agent (e.g., operating system and browser of requesting client), and result (e.g., error code, number of bytes sent to client). Other information such as referring URL and cookie codes may also be included. Server software can be configured to record every request, filter out requests for graphics or style sheets, or filter on other conditions such as error codes. An example from the BLS logs is shown in Figure 2. In all four lines, the time, IP address (anonymized in that the IP address is fictitious), requested page (immediately after the GET), error code (200 is a normal return, 304 is a redirect to an updated page), number of bytes sent to the client, client browser, system platform (e.g., Windows), and referrer page (page from which the link that made this request was called). Note that in lines one and two, the user went from (was referred from) one BLS web page to another. In line three, the user did a search using the Occupational Outlook search tool (for marine biologist). In line four, the search request came from the Lycos search engine. Thus, each request can take from a few dozen to hundreds of bytes and there are an enormous variety of URLs, search terms, and other parameters. The resulting logs grow quite large for popular websites and storage, backup, and maintenance are non-trivial. In sum, it is important to note that log analysis is constrained by the log settings on the web server—different servers may keep more or less data for each request.

Figure 2. Four sample lines from raw BLS logs

1. 09:22:24 902.00.00.00 - GET /sahome.html - 200 7884 Mozilla/4.0+(compatible;+MSIE+4.01;+Windows+98) - http://stats.bls.gov/datahome.htm
2. 09:22:25 800.00.00.000 - GET /oco/ocos249.htm - 200 108373 Mozilla/4.0+(compatible;+MSIE+5.5;+Windows+98) - http://stats.bls.gov/oco/ocoimo.htm
3. 02:09:23 100.000.0.000 - GET /oco/style.css - 304 141 Mozilla/4.0+(compatible;+MSIE+5.01;+Windows+98) - http://stats.bls.gov/aspsrch/oco2.idq?CiScope=%2Foco&IDQFile=%2Faspsrch%2Foco2.idq&SearchArea=%2Foco&CiSearch=marine+biologist
4. 20:09:24 600.000.000.0 - GET /oco/ocos122.htm - 200 35610 Mozilla/4.75+[en]+(WinNT;+U) - http://hotbot.lycos.com/?MT=effective+communication+for+financial+planners&II=10&RPN=2&SQ=1&TR=21358

Second, once the server creates the logs, a variety of procedures and tools are available for managing and analyzing them. Web log analysis software can be included in the server software (e.g., providing some minimal accounting of how many requests), but most large-scale websites use special software for analyzing logs. Example commercial analysis tools include: NetTracker; Microsoft Usage Analyst; and Web Trends Log Analyzer. There are also public domain analysis tools, notably, analog 5.0 http://www.analog.cx) and wwwstat (http://www.ics.uci.edu/pub/websoft/wwwstat/). In addition to the analysis packages, interesting commercial tools and techniques for mining and visualizing these data have appeared (e.g., see Eick, 2001). Since October 1999 (since the conversion to the NT server), BLS used versions of Web Trends for log analysis and it used Microsoft Usage Analyst in 1998. Previous to the NT conversion, Labstat staff wrote custom programs to summarize and analyze web logs that were created by the previous Unix web server. Results from these two different server log processes and log analysis programs were used in transaction log analysis over the years. In addition, we used a variety of customized (C programs) and commercial tools to conduct log analyses in the 1996 and 1997 periods (see those reports for details). In the 2000 analysis, Perl scripts and Java programs were created, but BLS data summaries were used whenever possible.

Permutations and Limitations of Transaction Log Analysis as a Methodology

As can be see from the discussion above, depending on what is logged by a web server, many types of analyses are possible. Analyses of system performance might look at error codes (frequencies and distributions), number of bytes served, number of requests handled, hit counts for pages (to examine site architecture, assess use of specific services such as help or feedback, or evaluate content value), or other values in combination. Most log analysis is done to assess user behavior. Analyses of activity factors such as request periods (day, time), client settings (geographic place, types of platforms and browsers, and referrer (especially important in commercial environments where advertising is used) are of typical interest. Another user behavior that is of interest is querying—what are people searching for? How are they searching for it? What terms do they use? Examination of query strings aims to address these types of questions. Some analyses of this type limit queries to those generated by the internal site search engine and other use referred queries from public search engines such as Yahoo or Google. A good recent example of the variety in data that can be used in transaction log analyses is given by Chen and Cooper (2001)--they clustered data from 47 search variables to find six categories of user behavior in web-based online public access catalogs.

Another factor that is of natural interest is the concept of session—how long do people stay at a site? A page? How many pages do they view in a session? Since the HTTP protocol is a 'stateless' protocol (the client's request goes to the server and as soon as it is filled, the connection between client and server is broken), considerable effort goes into defining and determining sessions. Some servers now artificially maintain the connection, or require log ins that maintain state for that user, and many send 'cookies' (unique identification strings that are stored on the client's browser) for this purpose. In government websites, where privacy issues are taken seriously, there are typically no explicit states kept nor cookies sent—this is the case at BLS. Without some record of state, analyses of transaction logs must make inferences about activity to identify sessions. For example, the IP address is treated as a single individual and some time interval used to aggregate all requests from that IP in that interval into a session. The assumption about an IP address representing an individual is mitigated by shared workstations in labs and by dynamic address assignment by Internet Service Providers.

A crucial session length factor is the amount of time a user is idle before considering a session to be terminated. Short idle time assumptions are tuned to casual users who come to a site and

browse or search for specific information, but suffer from counting lengthy sessions with lots of reading or interruptions as multiple sessions. Long idle times are tuned to more 'power' users who do lengthy reading on a single page or stay connected to a site over long periods of time that may include work interruptions, but may suffer from counting different sessions as one due to dynamic IP addresses or shared computers. Many researchers use 30 minutes as a sensible idle time interval for a session delimiter (the Web Trends default setting is 30 minutes for this purpose). In our 1996 and 1997 analyses at BLS, we assumed 60 minutes. In the customized analyses for the October 2000 logs we used a 10 minute idle time delimiter.

Finally, analysts might wish to study users' patterns of behavior—the sequences of actions they take during a session. These analyses quickly lead to combinatorial explosion as the number of possible paths in a site of thousands of pages (nodes) is known to be computationally challenging (algorithms do not run in polynomial time, but slow exponentially as more data is added linearly). Selecting a small number of specific paths to study is a prudent alternative to a generalized solution, but the nature of web page naming demands that customized parsing scripts be created for each website or these paths be specified in analysis profiles in tools like Web Trends.

It is important to keep in mind that because the log formats are complex, even some of these simple analyses might require custom scripts to be developed, run, and tested. Additionally, it is important to note that log files tend to be very large--a simple sort on one field in a file of several million records might take tens of minutes on a powerful workstation. Thus the many permutations in what is logged and what kinds of analyses are desired combine to offer a significant cost to transaction logs beyond the most basic and common summaries. In sum, the nature of web logs makes transaction log analyses complex undertakings.

As if the permutational problem were not challenge enough, there are some very significant inherent limitations to log analysis. Foremost, is the issue of page caching. Web browsing software helps save the user from reloading pages they wish to see multiple times or must see as part of a navigation pattern like backtracking. The browsers do this by maintaining the results from a page request in local memory for some time specified by the user in their browser preference settings. Thus, a user of a web site may go to a site's main 'homepage' and then to many other pages with several returns to the homepage along the way. However, because the homepage is cached on the client, there is no subsequent request sent to the web server and thus, no additional entry in the web server log file. This of course saves time and effort for both the user and the website, but does have serious undercounting effects on the representativeness of the web logs. A study done as part of our previous work (Fieber, 1998) suggests that the homepage of the BLS website may be underestimated in the transaction logs by a factor of 10. Less-frequently accessed pages are far less undercounted due to caching. This seems to be an open problem in transaction log analysis methodology at this time.

In spite of these limitations and complexities, web log analysis provides valuable views on system performance and user behavior. The overarching assumption is that as long as each of the parameter decisions are sensibly made and described in interpretations, that the large volume of data will overwhelm errors and yield realistic gross patterns. Properly qualified web log analyses can be powerful adjuncts to other data when assessing website use and planning for ongoing development.

Approach Taken to Transaction Log Analyses.
As the goal in this work was to look at trends over time, several types of data were gathered and analyzed. First, BLS provided a set of summary reports produced by LABSTAT personnel. The most extensive report is the Annual Report on LABSTAT Public Access Usage Statistics

January-December 2000 produced by the Quality Control Staff in the Division of Data Dissemination (BLS, 2000b). Other reports included yearly summaries of overall usage from 1996-2000 and query summaries for search engines. These reports are based on a set of log filters (e.g., no .jpg, .gif, .bmp requests, only GET requests, no requests with error codes, no zero length sessions) and 30 minute idle time settings. Whenever possible, these data were used in analyses and comparisons to 1996 and 1997 results.

Second, BLS provided the raw transaction logs for the month of October 2000. As the October 1997 logs and a similar period in 1996 (part of September and October 1996) had been analyzed in earlier work, the aim was to look at some gross comparisons across these time periods. A variety of tools were used in managing and analyzing these logs, including application packages such as Excel and SPSS, a variety of Unix utilities and original shell scripts, and customized Perl scripts and Java programs. The raw logs (1.5 gigabytes, containing 8,656,502 requests), were contained in 31 separate files, one for each day of October 2000. A variety of counts were obtained for different types of requests in these raw logs. Next, the logs were stripped of requests for style sheets (css), .gifs, and .ico requests. Java scripts were written to parse the raw logs into 31 files (one for each day) of sessions. One important (and costly) addition was a Domain Name System (DNS) lookup for each IP address to find the DNS name and add this to the parsed record. The parsed data was structured as follows:

session ID;total # of events in session;domain name/IP;total session time;OS;browser;URL hit|time from start (multiple of these);query string, ("-" if it wasn't a query).

As these analyses were undertaken, differences in how data are collected and summarized became apparent. Even when the same simple measure is desired (e.g., number of requests to the server for a specific page), the results depend crucially on what is actually recorded in the logs and of the data recorded, which values are included as candidates, i.e., how the logs are filtered before analysis. For example, whether a log entry is included for a mistyped URL or not, and if it is included in the log itself, whether it is a candidate for counting because it did not return any data. In the summaries below, these differences are discussed as they arise in the data presentation.

Transaction Log Results

The data are presented in two main sections—the first looks at growth in BLS website activity over the 1995-2000 period and the second looks at changes in user behavior as represented in the logs over this time. First, a brief summary of data from the 2000 log analysis is provided.

Summary of October 2000 logs. The October 2000 transaction logs were obtained from BLS and analyzed with an eye toward comparison with 1996 and 1997 analyses. The raw logs were contained in 31 files (one for each day). These files totaled 8,656,502 requests. Of these, 154,847 (1.8%) returned 404 error codes (page not found), usually due to mistyped URLs or out-of-date bookmarks, 16,454 returned 401 errors and 28,915 returned 403 errors (two types of requests to unauthorized pages). 132,260 302 codes and 9,583 301 codes redirected user requests (the 302 codes are typically transparent to the user), and 2,960 requests were not fulfilled due to server overload (502 code). Thus, about 98% of all requests returned a BLS webpage.

A total of 1,525,387 requests contained the string ‘search’ (either upper or lower case s), indicating one form of query. This number (17.6% of all raw requests) includes most requests from external search engines as well as most internal queries. Using the raw logs as a base, Google referred 136,140 queries to BLS in October 2000, with 117,774, 24,576, and 23,904 referred by Yahoo, Lycos, and AltaVista respectively. Note that Yahoo uses Google for query terms that do not match its category scheme so that some of the Google referrals actually

originated at Yahoo.

52,653 (0.6%) requests were for .gif files, 49,165 (0.6%) were for favicons[2], and 941,661 (10.9%) were for style sheets. To parse these raw logs into sessions, the requests for .gifs, .css, and favicons were removed. This left 6,096,158 requests. This compares with the 6,486,473 document views in the BLS report—within 6%). A Java program examined each request, did a domain name lookup for the IP number in the request, and collected requests from a single IP address with a ten-minute period threshold into session files. This parsing into sessions yielded at total of 1,432,304 sessions. This compares with 1,117,519 sessions in the BLS report—28% difference. This discrepancy is partly due to the relatively short session threshold (10 minutes) used for parsing sessions, and also to the differences in filtering used (BLS filters error codes and HEAD requests and known bad or missing pages). These distinctions highlight the difficulty of comparing transaction data over time as technologies and policies change. In the results that follow, the BLS Web data is used except in the noted cases where specialized analyses were added.

Activity Growth at BLS Website. The growth in Internet usage is well documented in the popular as well as academic press. The Pew Internet and American Life Project estimated that 104 million Americans had access to the Internet at the end of 2000 (www.pewinternet.org) and it is obvious that growth rates must slow as we reach saturation of the entire population. The usage of the BLS website also continues to show more usage over time. To look at this growth and assess whether it is approaching some stabilization, data for overall access was compared for the 1995 to 2000 period. Table 1 provides these data based on BLS internal reports and Figure 3 depicts this growth graphically.

Table 1. Total BLS Requests for Octobers 1995-2000

	Oct. 1995	Oct. 996	% inc	Oct. 1997	% inc	Oct. 1998	% inc	Oct. 1999	% inc	Oct. 2000	% inc
Total Requests	191,639	620,430	224%	1,490,328	140%	3,703,714	149%	4,761,160	29%	6,486,473	36%

October is a typically busy month for BLS (and other sites) and represents a good upper bound base for this snapshot of activity over time. Although the number of requests continues to grow, the huge increases in the early years—tripling volume, has decreased to less than doublings in the most recent years. To use a physics analogy, the velocity of requests continues to increase but the acceleration has begun to abate.

To put this rate of growth in perspective, the BLS monthly totals are juxtaposed with data from the Library of Congress (LC) website. The LC data (available on the LC website) is a count of all files transferred and thus includes images and other data not included in the BLS summaries. Additionally, the LC figures are given as yearly totals; the values here represent 1/12 of those values (since October is also one of the busiest months each year at LC) these LC values may be somewhat underestimated). Table 2 presents the data and respective increases over the six year period. Both BLS and LC showed more than triple the previous year's activity from 1995 to 1996. LC maintained this increase the following year, whereas BLS increased another two and a half times. From 1997 to 1998 BLS requests again increased another two and a half times while LC increased less than two times. Thereafter, increases at both sites have moderated very similarly. The Pearson correlation coefficient for the respective values over these years is 0.992. The important point is not to compare the actual values but the comparable growth in requests.

[2] A favicon is an icon that sites add as a logo, using an .ico extension. The IE 5 browser looks for this icon whenever users bookmark a site and then use it in the bookmark list. Presumably, almost 50,000 requests from IE 5 users were bookmarked in October 2001.

Figure 4 shows the respective rates of increase in these two sites over the five year period. Clearly, the early growth has moderated and both sites are no longer even close to doubling in activity each year. Another way to contextualize this growth is to look at overall growth in access to the Internet by the US population. The Pew survey data (Rainie & Packard, 2001) show an increase of 18% in Internet access in the second half of 2000 (from 88 million Americans in May-June to 104 million in November-December). Based on these data, it is reasonable to infer that one-third more Americans gained access to the Internet in 2000—a value that compares with the increase at BLS in 2000. Whether the growth of BLS activity is mainly due to these newcomers or to increased usage by past newcomers who have come to value BLS data more often is a question for future study.

Figure 3. Website requests at BLS in Octobers 1995-2000

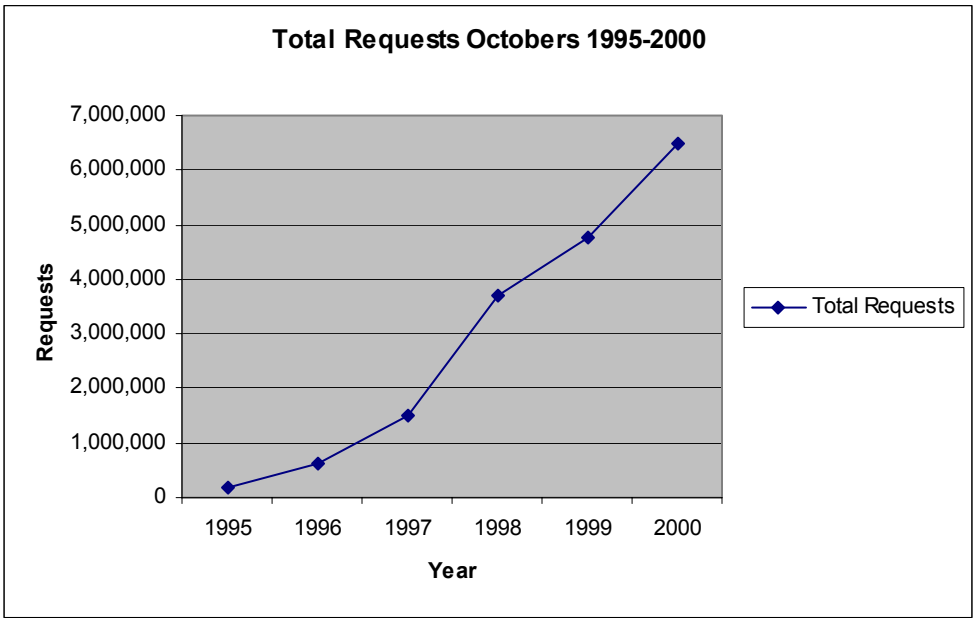
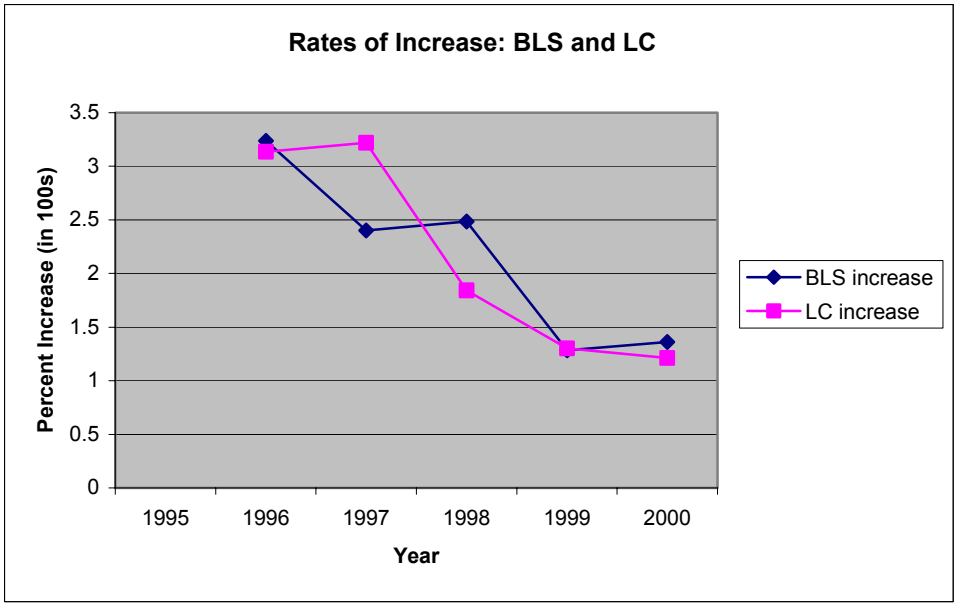


Table 2. BLS and Library of Congress monthly requests 1995-2000

	BLS	LC	BLS increase	LC increase
1995	191,639	1,981,045		
1996	620,430	6,214,470	224%	214%
1997	1,490,328	20,009,561	140%	222%
1998	3,703,714	36,876,884	149%	84%
1999	4,761,160	48,032,242	29%	30%
2000	6,486,473	58,268,221	36%	21%

Figure 4. Rates of Increase at BLS and LC



Characterizing User Behavior at BLS. Another way to look not only at the growth in BLS activity but also what types of services and data people are using over time is to examine what pages within the website were most often accessed over time (the BLS 2000 LABSTAT report provides detailed breakdowns on most-used pages and services). Table 3 presents data for these highly-requested pages. The data require qualification in several ways--logs were created by different servers and different types of analysis software were used across time; and the website changed as pages were renamed or replaced. The 1996 and 1997 data were obtained from BLS internal reports that were based on scripts written by BLS personnel to summarize the logs from the original UNIX-based server. The 1996 data covered the period September 24-October 23 and the 1997 data covered the period September 22-October 22. The 1998 data was based on a report generated using the Microsoft Web Analyst software, and the 1999 and 2000 data were based on reports generated using Web Trends software. It is also important to note that all of these top-level page names have variants and these data only reflect these specific requests. In the case of blshome, the data for 1996-1997 reflect blshome.html and 1998-2000 reflect blshome.htm. The eag (Economy at a Glance) page value for 2000 is for the eag.map page that replaced it. The oco/oco1000 Occupational Outlook Handbook index page in 1996-1999 was replaced by oco/ocoiab data for 2000. Search at BLS has perhaps undergone the most change in terms of how it is handled by the website and how the logs reflect user queries. In the 1996 and 1997 logs, there were entries in the logs for a request to present a query form. The queries actually entered by users were handled by cgi scripts and logged separately. The values in Table 3 reflect the number of times people clicked on the keyword request button. In subsequent years, queries were handled by specialized search engines within BLS and the 1998-2000 data in the table reflect search.asp requests. In addition to the general search facility available on the home page, there are specialized search functions within the BLS website. For example, the Selective Access service allows users to develop sophisticated searches using customized forms. The Occupational Outlook Handbook also has a search facility within that publication. In 1998 and 1999, the search/oco requests were easily among the most requested pages/functions and are included in the table to show this volume. The 2000 summaries did not include these requests separately so the 294,294 value reflects the total number of requests that contain search/oco in the raw logs and thus cannot be directly compared to the1998 and 1999 values that were based on counts after

filtering the logs.

Table 3. Top-level Page Requests for Octobers 1996-2000
(see text for notes on italics)

	1996	*1997*	1998	1999	2000
Total Requests	620,430	1,490,328	3,703,714	4,761,160	6,486,473
Top-level pages					
blshome	80,604	135,028	195,921	268,137	367,704
cpihome	12,957	33,818	47,367	62,520	88,382
dathome	42,866	63,613	92,411	116,360	128,163
eag	16,761	28,411	39,917	57,736	66,195
oco/oco1000	11,578	40,648	54,280	60,434	122,329
ocohome	31,440	93,269	152,493	207,690	309,342
proghome	11,207	15,673	27,400	32,415	32,539
sahome	14,408	20,153	26,174	31,101	33,460
top20	21,988	33,694	46,361	59,700	66,483
search	12,127	24,593	56,022	66,790	66,049
search/oco			138,999	198,440	*294,294*

Keeping these different qualifications in mind, it is easy to see that the growth in the BLS website usage has been distributed over the same set of pages over the five years 1996-2000. This is due to a combination of function/content and site architecture. In the former case, these pages are what people come to BLS to find. In the latter case, their popularity is due to their position in the BLS site—especially on the home page. Three exceptions to the home page positioning are the two pages from the Occupational Outlook Handbook and the CPI homepage. Use of the Handbook continues to dwarf all page requests with the exception of the BLS homepage. The importance of the CPI is reflected by the fact that it is a top-requested page and continues to increase in popularity over time. These trends have been taken into consideration in the BLS redesign released in November of 2000 in that the latest CPI values are available on the home page (dynamically generated) along with a link to the details and the Occupational Outlook Handbook is now available from the home page rather than at lower levels in the site architecture. These specific examples demonstrate the more general relationship between a website's structure and usage; and how transaction logs may be helpful in refining the structure of a site.

Table 4 shows the percentage increases for the top-level pages over the five year period. The data show that there was a ten-fold increase in overall requests to BLS over the period, but these increases were not proportionally distributed to the different top-level pages. Only the Occupational Outlook Handbook pages have shown parallel growth over this period. This suggests that much of the growth in usage is coming from students in schools (traditional users of OOC) or citizens exploring job opportunities. In either cases, these are non-specialist users of statistical data. The other top-level page that shows substantial increase in usage is the CPI home page, another page that has popular appeal among the non-specialist citizenry given the attention the media give and how pricing affects ordinary lives.

Table 4. Percentage Increases for Top-Level Pages 1996-2000

	1996	97 % inc	98 % inc	99 % inc	00 % inc	1996-2000 % inc
Total Requests	620,430	140%	149%	29%	36%	945%
Top-level pages						
blshome	80,604	68%	45%	37%	37%	356%
cpihome	12,957	161%	40%	32%	41%	582%
dathome	42,866	48%	45%	26%	10%	199%
eag	16,761	70%	40%	45%	15%	295%
oco/oco1000	11,578	251%	34%	11%	102%	957%
ocohome	31,440	197%	63%	36%	49%	884%
proghome	11,207	40%	75%	18%	0%	190%
sahome	14,408	40%	30%	19%	8%	132%
top20	21,988	53%	38%	29%	11%	202%
search	12,127	103%	128%	19%	-1%	445%

Another way to look at users' behavior is to examine how they search and what queries they use. It is difficult to accurately compare even gross patterns such as whether people are posing more analytical queries rather than browsing for data. An analysis of queries is beyond the scope of this work, but a cursory look at the most common queries sent to BLS via external search engines such as google.com suggests that there continue to be large numbers of queries on particular occupations covered in the Occupational Outlook Handbook and large numbers of queries for terms related to employment (e.g., career, job, employment, labor, occupation, unemployment, wages, compensation) and cost of living (e.g., CPI, costs, inflation).

Yet another view of user characteristics is to examine the settings in which they access the BLS website. Log analyses provide several aggregate portraits in this regard. Four types of data of interest are: top-level Internet domain from which the request arrives, type of browser and platform making the request, and referring website. Table 5 presents requests by top-level domain by session. The 'other' domain category was computed by subtracting the sum of the respective five top-level requests from the total number of sessions counted in each data set; The 1996 and 1997 data did not count .org. In our session parsing for 2000, we also counted the number of .com sessions that were from .aol (America Online) and found that half of the .com requests (222,804 or 49%) came through .aol, suggesting a very large home usage pattern. Over the five-year period, the .edu portion of requests declined in spite of traditional heavy access to the Occupational Outlook Handbook from schools. The data suggest increased home usage as more people gained Internet access at home through Internet Service Providers.

The October 2000 data show that Microsoft's Internet Explorer (IE) and one of the various Window platforms (95/98/00/NT) dominate usage with IE used for 66% of the requests the Windows platforms used for 91% of the requests. Although the data for browsers were not analyzed in the 1996 and 1997 reports, it is surely the case that IE replaced Netscape as the browser of choice and Windows-based platforms have solidified their position in the marketplace.

Table 5. Requests by Top-Level Internet Domains

Domain	Oct-96		Oct-97		Oct-00	
com	29858	17%	73904	25%	464853	32%
edu	19025	11%	41382	14%	135437	9%
gov	2224	1%	3743	1%	13865	1%
net	17329	10%	45629	16%	338538	24%
org					25534	2%
other	102588	60%	129105	44%	454077	32%
Total	171024		293763		1432304	

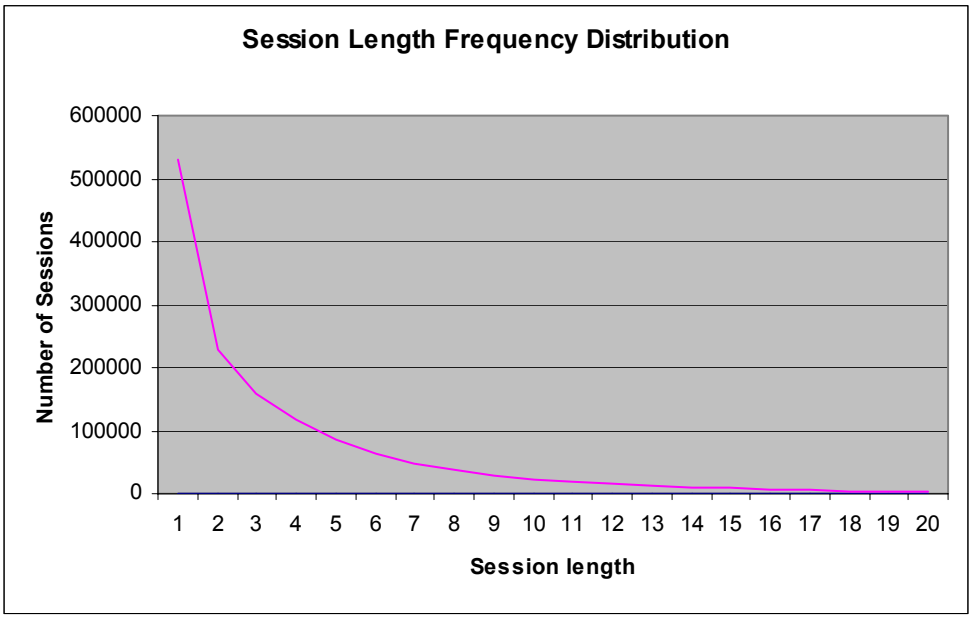
Referrer data t is useful in determining which external websites users were using immediately before arriving at BLS. In October of 2000, by far, the largest number of the requests with a referral came from search engines. Various Yahoo pages led the referrals (the Yahoo directory, google searches from Yahoo, and search.Yahoo directed 64631 requests in October 2000), with Microsoft-NBC search (43,737), Google (24,292), and AltaVista (11,082) also referring tens of thousands of searches to BLS. The only non-search engine referring site with more than 10,000 referrals in October 2000 was the Department of Labor main page (.dol.gov). Clearly, many people arrive at BLS through searches posed to popular search engines and the query analysis discussed above suggests that these people are typically looking for career information or basic information on cost of living or employment conditions—i.e., casual users who are not specialists in statistics. Of course, we have no way of knowing how many user queries should have been referred to BLS but were not (either through poor query formulation on the part of the information seeker or through poor indexing on the part of the query engine).

That most BLS website users are non-specialists is strongly reinforced by the percent of users who visit the site only once in the period of analysis. For the month of October 2000, 458,679 unique users visiting the website and 82% of these visitors came to BLS exactly one time. Almost 10% of the visitors made two visits in the month of October. Less than 2% of the users visited the BLS website ten or more times in a month. It is instructive to keep in mind that although 2% is a small proportion, it does represent more than 9000 people who visit the site quite often. Keeping in mind the large volume of non-specialist as well as these intensive users when planning for services is a difficult balance to achieve. As with the other data, these data are qualified by the fact that a person could use a different machine with a different IP address and be counted as a separate individual; likewise, someone coming from the same home machine through a ISP that dynamically assigns IP addresses will be likely counted as a different user each time and multiple users from different places could be counted as the same user through the dynamic assignment. Thus, these numbers and percentages should be taken as upper bounds since the same user might get counted as a different user each time they use their ISP.

Another indicator of the non-specialist user is the pattern of usage once people make a request to the BLS website. In the 1996 and 1997 reports, as many as half of the sessions were of length one---that is people only requested one page and made no further requests within the one-hour idle time threshold. The BLS data show that in October 2000 there were 145,684 one-request sessions (13% of all sessions). Our analysis shows that 37% (529869 of 1,432,304) of all sessions were of length one. This difference reflects the fact that our analysis did far less filtering and used a shorter idle time cutoff. Our data show that 95% of all sessions are completed in twelve or fewer requests (37%, 16%, 11%, 8%, 6%, 4%, 3%, 3%, 2%, 2%, 1%, and 1% respectively for the length of session 1-12 requests). This data shows the classic hyperbolic

(Bradford) distribution shape commonly found in library circulation and other usage behavior patterns. Figure 5 depicts this distribution for our October 2000 data.

Figure 5. October 2000 Session Request Length Distribution



These different views of the log data give glimpses into the gross behavior of users of the BLS website. In summary, the BLS transaction log data show:

- Continued growth in usage since the introduction of the BLS website, but a decreasing acceleration in growth;
- Most users are non-specialist, casual users who visit the site only occasionally and then for relatively short periods of time, and often access BLS from home;
- Technological migration and consolidation toward Wintel platforms and software;
- Usage is strongly related to site structure.

The results also demonstrate how transaction log analysis can be helpful as one indicator of user behavior and to suggest directions for website maintenance and improvement. They also demonstrate the limitations of transaction logs alone, the many parameters that must be considered with this technique, and the difficulties in making comparisons as technology and policies change.

Framework for Organizational Interface Evolution

Two themes emerge from the interview and transaction log data: the BLS website and its user interface has become a significant element in the BLS organizational interface, and this change has led to issues and changes in all the dimensions of the organizational interface—to wit, the user and organizational interfaces have co-evolved. In general, the five years of website usage from 1995-2000 indicate that the Internet services have become part of the BLS infrastructure rather than an add-on to the data dissemination mission of the institution. The Internet and Intranet have become key elements of BLS' organizational interface with the respective user interfaces to the public website and Intranet as entry points. This theme is evidenced in several ways. First, the interviews show that both BLS employees and the public have come to depend

on the BLS website and other Internet services like email, and continue to have additional expectations as a result of this dependence. This dependence is also strongly shown by the transaction logs in more moderate but still strong growth in usage over time. It is unimaginable what might occur at BLS if the website were to disappear.

Second, BLS has devoted significant resources to the development and maintenance of the website. BLS personnel respond to an increasing number of email requests that in many cases arise after or while people use the website. Furthermore, the LABSTAT staff respond to a large number of requests specific to the website. A number of new services have been added over the years. The webserver itself is now a commercial enterprise—moving from an early Unix environment retrofitted to web use to the dedicated NT environment used today. Likewise, commercial products are used in maintenance and development.

Likewise, the development team has a long-term plan for upgrades and the user interface underwent several minor revisions from 1996-2000. The website undergoes usability testing and an iterative approach to maintenance and redesign have been put in place. Our interface prototypes, vocabulary analyses, and usage assessments have influenced the BLS website substantially through structure reorganization (e.g., minimizing clicks to popular services and data), use of non-specialist-oriented terminology, and user-centered access mechanisms (e.g., at a glance summaries, maps, and tooltips). In 2001, a major interface revision was made and the website is now significantly different than during the period studied in this report. It will be interesting to see how this major re-design affects some of the long-standing activities of users.

In addition, new types of services have been added to the website over time. Special pages for use in K-12 environments were added, specialized summaries for non-specialists like economy-at-a-glace were introduced and refined, new tools such as an inflation calculator were created, email alerting services were instituted, new document formats were included (pdf), and access to additional databases was provided. All these new services provide better access to the public but also demand resources and the fact that BLS continues to add such services demonstrates the commitment that upper management is making to Internet-based dissemination.

Third, in addition to these added services, there is a level of maturity in the way that information and the website are viewed and managed. Security of data and the timing of data releases are important themes that garner the attention of everyone from senior managers to division staff. The in-house usability testing and iterative design process is another indicator of infrastructural maturity. This maturity is reinforced by the institutionalization of internal reports for transaction logs and email usage. The compilation of the "Annual Report on LABSTAT Public Access Usage Statistics January-December 2000" is a strong indicator of how BLS has begun to value transaction log data and use it in planning and developing future iterations of the BLS public access services.

The second theme relates to issues arising from the institutionalization of the website and other Internet services and their user interfaces. These issues reflect the interactions of technology, data, people, and the organization and are manifested in the current state of the BLS organizational interface. One thread of this theme relates to how resources will be allocated to serve the growing number of requests that easy access causes. Easy access encourages larger, more diverse customer groups, who in turn require help with understanding and using data, statistical methods, and technology. The demands range from increasing volumes of phone and email requests to efforts to make the website more user-friendly and appropriate to diverse needs and experiences.

Another thread in this theme relates to issues of mission and service. The media attention to and strong influence of BLS data on markets raises awareness and scrutiny among the public as well as Congressional leaders. In the past, the enormous efforts made to collect and analyze labor-related data were of concern to a small subset of the population who leveraged these data for various specialized purposes. As larger portions of the populace seek and use BLS data, more types of data are requested, implying pressure for more resources and efforts on the part of BLS. More importantly, more attention and more questions will obtain—requiring more reflection and quality control than ever. This was reflected in the concerns in the interviews about levels of review for reports as well as email responses to the public. These concerns are likely only the tip of the iceberg in terms of new demands for data and for quality assurance in whatever data is provided to the public.

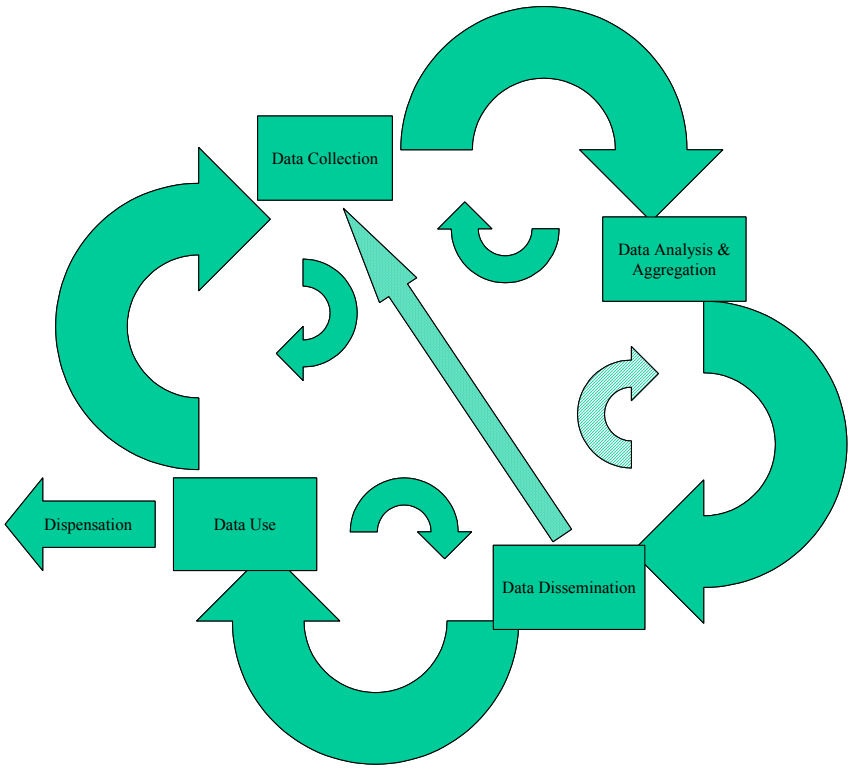
Ultimately, it is likely that the information life cycle will be affected at BLS. The traditional life cycle has been discussed by many researchers over the years (e.g., Levitan, 1982; Taylor, 1982; Hodge, 2000). A version of the cycle is represented in Figure 6. In this scheme, data is collected, analyzed and aggregated, disseminated, and eventually used by consumers or may be disposed of in archival setting. Data usage in turn may affect what new data is collected, thus completing the cycle. In each of these phases, experience suggests that there are feedback channels where each phase influences its predecessor. We are beginning to see ways in which BLS' dissemination effort via the website has begun to influence data analysis and aggregation through new releases, new summaries aimed at non-specialists, and new tools for meeting diverse needs. Thus, the feedback arrow that is hatched in the figure is strongly supported by the data in this study. Likewise, how citizens, journalists, and others use the data may influence what data is disseminated—certainly we see evidence for users influencing the form of dissemination (e.g., pdf files as well as HTML and ACSII). The interfaces used in the dissemination phase may in fact eventually propagate back two steps to what data is collected—if easy access causes more people to request additional data not currently provided, Congress or BLS leaders may adapt existing surveys or create new surveys to meet these needs. This hypothesized feedback is represented in the figure by the dotted arrow from data dissemination to data collection. Documenting such trends must be done over long periods and this case study is a beginning toward this goal.

To better understand these changes, especially how feedback channels and growing usage affect the life cycle, the following framework for organizational interface change emerges from the five years of data. As noted in the introduction, technology (hardware and software), data, people, and organizations are dimensions that interact to determine changes in human and organizational behavior and the organizational interface manifested by these behaviors. To these main dimensions, two additional dimensions are introduced that emerge from interactions among the primary dimensions. These additional dimensions are: interaction style and population literacies. Figure 7 summarizes these dimensions with respect to change over time and along the respective subdimensions of internal and external user perspectives. The changes in hardware and software are typical of all organizations and users as advances have led to faster machines, better Internet connections, and more advanced graphical user interface browsers.

There are some changes in the types of data that BLS provides--mainly adding additional data sets and creating new aggregations that are 'born digital' to help non-specialists access often-requested data. From the external user's perspective, it would be useful to analyze whether the distribution of all page accesses has changed over time, but we are limited to top-level page access (which in turn are affected by caching), which over the five years have remained relatively stable proportional to the growing number of visitors. Whether this will continue after the major redesigns is worth assessing in future work. Certainly, BLS has added new data such as the Kid's

Page and many new aggregations (e.g., Economy at a Glance summaries) and specialized tools (e.g., inflation calculator).

Figure 6. Information Life Cycle with Feedback Channels



Clearly, one large change is the continued variability in the user population as more diverse people become connected and use BLS. The growth in email requests, diversity of questions that come in, and number of visitors to the website all demonstrate this trend. The requests for Spanish versions of documents and requests for data that BLS does not have (and is out of scope—e.g., birth certificate) and does have but is not generally available (e.g., historical data, selected microdata) suggest that the biggest changes to BLS may be yet to come with respect to what data is collected and disseminated--the very core of the BLS mission. Quality control and perceived accuracy are crucial in this regard.

The interaction styles of internal staff reflect the global changes taking place in knowledge-intensive workplaces—more email communication rather than telephone communication, more exchange with people outside one's department or the entire organization, and in many cases, a sense of efficiency and higher productivity (with higher stress levels in some cases). The interaction styles of end users are becoming more dynamic---beyond the simple form fill in query and drill-down selection strategies of interaction, people are using tool tips and layers of information more naturally. BLS, like all government agencies has been careful to address universal access issues by providing text only alternatives and not adopting dynamic pages, flash pages, and Java applets, in spite of larger portions of the user population being ready to accept these techniques. The growth patterns show that the saturation of universal access is still a long way off, although more than half the population now uses the Internet. Because there are still many new people coming to the Internet, BLS will have to continue to provide a range of

Figure 7. Framework for Organizational Interface Evolution at BLS

Dimension	1995	2000
Hardware		
Internal	Various existing mainframe systems	NT-based dedicated servers
External	Early Wintel, Apple, PCs Slow dial up, T1 in orgs	Windows-based PCs fast modems, cable modems, ASL, fiber, wireless
Software		
Internal	Unix, C programs	Microsoft Products (server, analyst) Web Trends, other commercial
External	Lynx, FTP, gopher, Mosaic	Internet Explorer, others
Data		
Internal	Surveys, news releases, etc.	same, plus economy at a glance, Kid's Page, aggregations.
External	download what is available	clarifications, new requests, dynamic generation of results Multiple formats (paper, HTML, pdf)
People		
Internal	agency-oriented	customer-oriented, high-tech users personal accountability
External	specialists	specialists plus diverse non-specialists
Interaction Styles		
Internal	paper, phone, and face-to-face based	e-based
External	text based, drill down selection	search+browse across sites, short sessions, ready to email
Literacy		
Technical	small portion of population specialists adapt to technology	majority of population high expectations for ease of use
Statistical	small portion of population	small portion of population
Organization		
Internal	Agency centered	Customer centered High growth Quality control e-based infrastructure new policies
External	monolithic agency	responsive service agency

interaction styles to accommodate the user experience as well as the range of platforms. One change reflected in the evolution of the website user interface over time, and especially in the recent redesign based on this work reflects a data and user-centered style. This design is less graphical with much more data immediately available on the front page and organized in a user-oriented manner rather than agency perspective. To minimize clicks and speed data access, this design brings often-requested data and links to the surface from pages that were buried several levels into the hierarchy, centralizes often-requested indexes and values on the home page, and provides a clickable map for localizing data. This design makes the home page very information

intensive and business-like rather than requiring users to adopt a drill-down interaction style.

The overall literacy of the population is determined by combinations of technology trends as well as the proportion of adopters in the population and intensity of use by these adopters. Rogers' diffusion of innovation theory (Rogers 1995) includes stages of adoption in the population (early adopters, early majority, late majority, laggards) as well as characteristics of innovation (relative advantage, compatibility, complexity, trialability, and observability) that apply in the BLS website case. Norman (1998) has applied this model to interface design and it surely applies in the case at hand. Over the past five years we see evidence that by 2000 the BLS website had been adopted by the early majority as an advantageous innovation. However, it appears that people's behavior was not facile at this point (e.g., many abandonments, short sessions, naïve questions via email) and perhaps not all of the characteristics of innovation had been satisfied. Perhaps the data itself is too complex, or the statistical literacy is so poor that compatibility cannot be obtained. In either case, non-specialist users may have difficulty assessing trialability or observability of results. Surely, the technical literacy of the population increased over the five years, with the popular media illustrating the uses of technology to the point that computers, the Internet, and the WWW are part of popular culture. Although this popularization causes more people to access the BLS website, full adoption of the innovation requires additional progress on the other factors related to adoption, especially statistical literacy. It is clear that Internet adoption is quite far along the adoption curve with well over 50% penetration in most developed countries. However, adoption of digital statistical data is far less developed as people must learn the value of statistics for personal decision making as well as how to find, interpret, and use these data. Related work on electronic tables, including several types of user studies, demonstrates the many difficulties statistical literacy poses to even technically sophisticated users (Marchionini et al, 2001; Marchionini & Mu, in review).

BLS as an institution and its organizational interface has likely changed more than is apparent due to the adoption of the Internet and WWW-based user interfaces. The physical changes are highlighted in the other dimensions as new hardware, personnel, and budgets become institutionalized. The corporate culture changes are more subtle and are reflected somewhat in the commentary of BLS staff about changed work behavior, personal accountability, and morale. More basic changes in corporate policies and mission will take longer and some early indicators are evident in the security awareness and quality control levels. A significant change is that the website and its user interface has become part of the fundamental organizational interface of BLS and these interfaces will continue to interact with changes in technology, data, and people to influence BLS in the years ahead.

This framework can be used to look at the life cycle and predict what feedback channels are most active and where planning and resources can achieve most good. The BLS has in many ways become reflective about the long term implications of IT and is well-positioned to continue to plan based on systematic assessments of progress and reflections on process. This is manifested in LABSTAT growth within BLS, the institutionalization of a usability testing laboratory, and procedures for assessing usage and planning for new user interfaces.

In sum, BLS has adopted Internet technology over the past half decade and this adoption has in turn become part of the infrastructure of dissemination and organizational interface for the institution. This is particularly appropriate at an agency that is knowledge-intensive and driven by information collection, analysis, and dissemination. The BLS website and its interface began as an quasi-experiment and grew into a high-profile element of the BLS institution. The genie is out of the bottle and creativity and logic will be needed in the years ahead to continue to plan and manage not only the specific user interfaces people will use but the much more basic demands

and implications these services bring to the organization and how they affect the larger organizational interface that defines the institution.

To achieve the dream of universal access to the information and communication channels necessary to live and prosper in the twenty-first century, we must find ways to bring together data, people, technology, and organizations. The data systems, policies and procedures, corporate culture, and public face of an agency or institution make up its organizational interface. To this end, a general theoretical goal of this work was to develop a model of the co-evolution of user interfaces and organizational interfaces in statistical government agencies. This work added new insights into user needs and behaviors, and new principles and practices for interface design. It is especially clear that changing the user population (by admitting a more diverse group of capabilities and needs) changes all the other aspects of the data collection and dissemination enterprise. What is evident from this work is that electronic dissemination of statistical information and user interfaces devoted to non-specialists not only lead to more usage by non-specialists but also begins to change the data provider itself. This is a longitudinal and theoretical issue with implications beyond government agencies. Beyond its use as a theoretical beginning for a theory of co-evolution of interfaces, the framework arising from this case will lead to improved user interfaces, promote universal access as broader user populations take advantage of data, and help agencies respond to resulting changes and plan future services.

Acknowledgements:
This work was supported by contracts from the Bureau of Labor Statistics. Stephan Greene wrote the C programs and Ben Brunk wrote the Java programs to parse the transaction logs in the first and fifth years respectively. Anita Komlodi conducted usability studies for the interface prototypes. The author thanks the BLS staff who agreed to be interviewed for this work and Michael Levi, Demetrio Scopelliti, and Kate Donahue who provided data and reports on the BLS website; Mary Michael and Deborah Klein who provided data on BLS dissemination; and Fred Conrad, Cathy Dippo, and John Bosely, who provided feedback on the overall data collection effort. Carol Hert and Stephanie Haas conducted parallel work that informed this case study and both made helpful comments on a draft of this paper.

References

Barreau, D. (1997). Information systems for organizations and the problem of ephemeral information. Unpublished doctoral dissertation (University of Maryland, College Park).

Bickerton, D. (2000). Corporate reputation versus corporate branding: The realist debate. *Corporate Communications*, 5(1), 42-48.

Bureau of Labor Statistics. (2000a). *Bureau of Labor Statistics 2000 Customer Service Guide.*

Bureau of Labor Statistics Quality Control Staff: Division of Data Dissemination. (2000b). *Annual report on LABSTAT Public Access Usage Statistics: January-December 2000*.

Campagnoni, F. & Ehrlich, K. (1989). Information Retrieval Using a Hypertext-Based Help System. ACM *Transactions on Information Systems*, 7(3), 271-291.

Chen, H. & Cooper, M. (2001). Using clustering techniques to detect usage patterns in a web-

based information system. *Journal of the American Society for Information Science*. 52(11), 888-904.

Eick, S. (2001). Visualizing online activity. *Communications of the ACM*, 44(8), 45-50.

Fieber, J. (1998). Browser caching and web log analysis. http://fallout.campusview.indiana.edu/~jfieber/papers/bcwla

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.

Haas, S. (2001). From words to concepts to queries: Helping users find series and variables to satisfy their information needs. Final report to BLS http://ils.unc.edu/~stephani/bls/fin-rept-01.pdf

Hearst, M. (1999). User interfaces and visualization. In R. Baeza-Yates & B. Ribeiro-Neto (Eds.) Modern information retrieval. pp. 257-323. NY : ACM Press and Addison-Wesley.

Hert, C. (1999). 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.

Hert, C. & Marchionini, G. (1998). Information seeking behavior on statistical websites: Theoretical and design implications. *Proceedings of the 61st Annual Meeting of the American Society for Information Science*. (Pittsburgh, PA, Oct. 25-29, 1998). P 303-314.

Hert, C. & Marchionini, G. (1997) Seeking Statistical Information in Federal Websites: Users, Tasks, Strategies, and Design Recommendations. Report to BLS (summer 1997) http://ils.unc.edu/~march/blsreport/mainbls.html

Hodge, G. (2000). Best practices for digital archiving: An information life cycle approach. D-Lib Magazine, 6(1), January 2000. http://www.dlib.org/dlib/january00/01hodge.html

Levi. M. (2001). *The implications of Section 508 on BLS data dissemination*. BLS working report.

Levitan, K. (1982). Information resources as "goods" in the life cycle of information production. *Journal of the American Society for Information Science*, 33(1), 44-54.

Malone, T. (1988). Computer support for organizations: Toward an organizational science. In J. Caroll (Ed.) Interfacing thought: Cognitive aspects of human-computer interaction. P. 294-324. Cambridge, MA: MIT Press.

Marchionini, G. (1989). Information-seeking strategies of novices using a full-text electronic encyclopedia. *Journal of the American Society for Information Science*, 40(1), 54-66.

Marchionini, G. (1998). Advanced Interface Designs for the BLS Website: Final Report to the Bureau of Labor Statistics. Report to BLS (summer 1998).

http://ils.unc.edu/~march/blsreport98/final_report.html

Marchionini, G. (1999). An Alternative Site Map Tool for the Fedstats Website Report to BLS (summer 1999). http://ils.unc.edu/~march/bls_final_report99.pdf

Marchionini, G. (2000). Interfaces to support customized views and manipulation of statistical data. *The second international conference on establishment surveys (Buffalo, NY, June 17-21, 2000*). Alexandria, VA: American Statistical Association. 953-959.

Marchionini, G. (2000). From Overviews to Previews to Answers: Integrated Interfaces for Federal Statistics. Report to BLS (summer 2000). http://ils.unc.edu/~march/bls_final_report_99-00.pdf

Marchionini, G. & Komlodi, A. (1999). Design of interfaces for information seeking. In M. Williams (Ed.). *Annual Review of Information Science and Technology*. Volume 33. Medford, NJ: Information Today. 89-130

Marchionini, G., Brunk, B., Komlodi, A., Conrad, F., & Bosley, J. (2000). Look Before You Click: A Relation Browser for Federal Statistics Websites. *Proceedings of the Annual Meeting of the American Society for Information Science* (Chicago,, Nov. 12-16, 2000), 392-402.

Marchionini, G., Hert, C., Shneiderman, B., & Liddy, L. (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

Marchionini G. & Mu, X. (in review). User studies informing E-Table Interfaces. *Information Processing & Management.*

Norman, D. (1998). The invisible computer. Cambridge, MA: MIT Press.

Penniman, D. (1975). *Rhythms of dialogue in human-computer conversation*. Unpublished PhD Dissertation, Ohio State University.

Rainie, L. & Packel, D. (2001) more online, doing more. Pew Internet and American Life Project. http://www.pewinternet.org/reports/pdfs/PIP_Changing_Population.pdf

Rice, R., & Borgman, C. (1983). The use of computer-monitored data in information science and communication research. *Journal of the American Society for Information Science*, 34(4), 247-257.

Rogers, E. (1995). *Diffusion of Innovations* (4th edition.). NY: Free Press.

Shneiderman, B. (1998). *Designing the user interface: Strategies for effective human-computer interaction* (3rd Ed.). Reading, MA: Addison-Wesley.

Taylor, R. (1982). Value-added processes in the information life cycle. *Journal of the American Society for Information Science.* 33(5). 341-346.

Tolle, J. & Hah, S. (1985). Online search patterns: NLM CATLINE database. *Journal of the American Society for Information Science*, 36(2), 82-93.