For the past four years, a team of information specialists has been working with the Bureau of Labor Statistics to improve outreach to constituent groups by means of the World Wide Web (WWW). In our preliminary work, 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 (http://ils.unc.edu/~march/blsreport/mainbls.html). In the second year, additional user needs analyses were done (Hert & Marchionini, 1998), transaction log analyses were replicated, an interface prototype that aimed to provide alternative entry points for different user needs and types was built and tested, and recommendations for short and long term design strategies were made (http://ils.unc.edu/~march/blsreport98/final_report.html). In the third year, the interface work extended these efforts to the Fedstats website that includes the bulk of statistics collected and disseminated by 70 federal government agencies. Fedstats is a portal service (also called a locator service or a metasite) meant to serve as a gateway to all U.S. federal government statistics (www.fedstats.gov). An interface prototype (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 with users and recommendations for revisions made (http://ils.unc.edu/~march/blsreport99/final.pdf). In the 1999-2000 phase of the project, 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 (http://ils.unc.edu/~march/blsreport00/final.pdf). The details of the designs and the usability tests may be found in these reports. A brief summary of the goals and usability testing results are given here.

It is clear that more and more statistical information is being made available to the public via the WWW. Many government agencies aim to improve citizen access and usage of government statistics by making these data available. Broadening access and use requires user interfaces that are easy to use and serve the needs of wide ranges of people. The desire to make statistical information more widely accessible via WWW interfaces raises several challenges to designers:

1. Universal access challenges. There is a critical need to develop alternative interface designs because human abilities and skills vary widely, we must provide access to people with a range of special physical and cognitive needs, numerical literacy skills tend to be low in the overall population, and there is a wide range of statistical needs that might be served by government statistics. In addition to the range of human capabilities and needs, technology continues to evolve and the range of platforms, software and connectivity grows. There is a limit to new requirements that government websites can put on users since the variance in citizen platforms continues to grow—there is a kind of installed base inertia that builds up as larger portions of the population acquires WWW access. Additionally, the cultural phenomenon associated with the “attention economy” tends to demand immediate and concise (often simplistic) responses to inquiries, regardless of how complex the appropriate results may be.

2. Retrieval challenges. Online information systems offer people possibilities to search properly indexed databases in highly systematic ways or to explore opportunistically. People may use analytical search strategies via queries and selections as well as explore/browse with general queries and selections. WWW interfaces for statistical data must provide appropriate interface tools to support this entire range of search strategies.

3. Display. Once potentially relevant data is identified, it must be displayed and manipulated. People want information alchemy—answers customized to their specific needs, regardless of the form and substance in the underlying data. There are two interface issues here: what to represent (e.g., overviews of large collections; specific tables or graphs with associated annotations, etc.) and control mechanisms for reading/understanding/manipulating these representations (e.g., scroll, jump, drag, zoom, morph).

4. Metadata. Retrieval and display depend on indexes for salient data characteristics. These data about data are known as metadata. Dynamic interfaces require consistent metadata to support different ways to slice and view
results. If the metadata exists, it must be organized and incorporated into the interface. This requires additional data transfer to the client side. If the metadata is not readily available, it must be created—a costly undertaking for government agencies.

5. Interface Management. Just as data must be managed from a life cycle perspective (from creation to dissemination to preservation and dispensation), so interfaces must be managed. Agencies must create iterative redesign schedules, conduct iterative usability testing, and maintain active versions of the interface for users with different capabilities. This requires changes in policies and culture as much as changes in technical operations.

With these challenges in mind, we have taken a user-centered and iterative approach to interface design and focused on two particular issues. The design approach is to:

- Understand human information needs and behavior by observing, interviewing, and working with users.
- Base design on user needs as well as data characteristics
- Create prototype user interfaces
- Test these prototypes with users, and
- Revise prototypes and recommend process models for implementation and maintenance

In the BLS work, we focused on two particular principles that we believe improve access and use of statistical information: a) offer users alternative entry points and data views; and b) minimize mouse clicks to get to pertinent data. Alternative views (the Agileview framework defines overviews, previews, reviews, shared views, and peripheral views; Marchionini et. al., in press) aim to address the challenge of diversity in the user population. Minimizing mouse clicks is based on the observation that clicking on a WWW link causes users to wait for information to load and also changes the window focus—both of which lead to cognitive interrupts. The basic idea is that a click is a radical act and techniques to look ahead before committing to a click are desirable.

To this end, we created a set of interfaces that were tested and revised. The first prototype was an alternative entry point for the BLS website and aimed to minimize clicks through information abundant design and use a graphical region metaphor for layout. Figure 1 displays this interface prototype. Note the spatial metaphor that presents three views of the data (data retrieval, publications, fast info) and minimizes mouse clicks by placing large amounts of highly structured information on a single page. In the fast info “region,” commonly accessed data is made immediately available (current statistics and news releases) and common scenarios are used as simple entry points.
Usability tests were conducted with nine subjects in the BLS cognitive laboratory. Subjects conducted searches for assigned tasks and verbal feedback as well as video records of hands, face, and screen were collected. The results showed that although subjects liked the alternative approaches to data access, they found the screen complex and that it tended to overload them with information choices. See [http://ils.unc.edu/~march/blsreport98/final_report.html](http://ils.unc.edu/~march/blsreport98/final_report.html) for details on this prototype and usability testing.

A second prototype was created to provide overview information for the 200 websites from the 70 agencies accessible through the Fedstats portal to US government statistics. The aims of the Relation Browser tool were to provide an overview of federal statistics on one screen, to provide look ahead without clicking (through mouseover mechanisms), and to provide enriched data attributes to add to understanding of the data in the websites before people committed to loading them. The prototype was built as a Java applet. The first prototype crossed topics (14
topics used in the Fedstats site) with data types (the forms of data available in the statistical sites) and provided filters for geographic location and time. Figure 2 depicts this prototype.

**Figure 2. Fedstats Relation Browser Prototype 1.**

![Fedstats Relation Browser Prototype 1](image)

Usability tests were conducted with nine subjects in the BLS cognitive lab. Subjects conducted assigned searches using the Fedstats site map tool as well as the Relation Browser. Verbal data as well as video data were again captured and analyzed. Although subjects liked the filters, they were unable to use them effectively and sometimes got zero hits by over-filtering. A number of screen display suggestions were also made. Based upon these results, the prototype was revised and a field test conducted. Figure 3 displays the revised Relation Browser prototype.
After receiving OMB clearance, the RB tool was provided on the Fedstats home page as an alternative (labeled as an experimental sitemap) for the Sept 1999-May 2000 period. Clicking on the button loaded the Java applet that included a button to provide feedback. The feedback button led to a voluntary questionnaire that contained 10 five-point Likert-scaled questions, one check list question, and two open-ended questions. Although transaction logs showed that the RB tool was used several thousands of times during this period, only 74 individuals completed the questionnaire. The results showed strong support for adding the tool as an option (50% strongly agreed it should be added). One-third of the respondents reported using the tool from home, thus suggesting that the goal of serving the general public was at least partially met.

Based on the field test, a third iteration of the Relation Browser was implemented and a usability test conducted. Figure 4 displays the final version of the Relation Browser. As in previous usability tests, nine subjects conducted assigned searches with the Fedstats site map tool and the Relation Browser at the BLS cognitive lab. Subjects expressed strong satisfaction with the way that data types, time data, and regions were linked to categories and claimed that the tool would save them time by helping them avoid going to websites that did not have data relevant to their needs. They also noted that some basic level of sophistication and/or training would be needed to take full
advantage of the tool and that the font sizes were difficult to see. Overall, there is good support for incorporating the Relation Browser as an optional overview tool for federal statistics. See (http://ils.unc.edu/~march/blsreport00/final.pdf) for details on the design evolution and usability testing.

Figure 4. Fedstats Relation Browser Prototype 3.

This work has demonstrated that dynamic interfaces can help people gain an overview of statistical data, better understand the range and type of data offered at websites, and save people time as they seek statistical data. Moving from prototype to operational system requires changes in several areas of management. First, interface techniques
and options that serve diverse user communities and capabilities must be adopted and integrated into the agency design culture. Second, user-centered, iterative design processes must be adopted that systematize updates and new releases and integrate usability testing throughout the design and production process. Third, the process by which data is captured and released must include procedures for assigning and maintaining metadata that supports dynamic interfaces. This includes possibly classifying data according to attributes across different agencies. For example, each of the 200 people responsible for the different statistical websites linked from Fedstats must specify (index) how their data fits into the various data type, region, and time attributes. In some cases, this is a simple notation, in others, original classification decisions are required, which may in turn require policy decisions. These policies and classifications must be standardized across all the agencies and therein lays an organizational challenge. This final challenge is perhaps the most difficult one, as institutional inertia and turf protection must be overcome.

Although the challenges are great, the possible rewards of improving citizen access are considerable. These may include: equitable access to valuable decision making information; satisfying work experiences; improved government service (and citizen perception and satisfaction); augmentation of the intellect for individuals empowered by access; augmentation of the social intellect as the general population makes better personal decisions; and generally improved performance and quality of life. David Moore’s comment in the International Statistical Review summarizes the reason we should endeavor to overcome the challenges of universal access and use of government statistics: “Democratization is driven in part by the quantization of society.”

References

See the WWW sites referenced in the text for full reports.


THE ROLE OF KNOWLEDGE REPRESENTATION IN MANAGING STATISTICAL INFORMATION

Stephanie W. Haas, University of North Carolina
School of Information and Library Science, CB 3360, 100 Manning Hall,
University of North Carolina, Chapel Hill, NC 27599, USA
stephani@ils.unc.edu

ABSTRACT

The importance of knowledge representation in understanding and managing complex information resources is gaining increasing recognition. A clear, unambiguous representation of concepts, rules, sources, ranges, of values, etc., is fundamental to the organization and presentation of information in an intelligible and useful way. Providing citizen access to statistical information is a particular challenge, in that users have a wide range of expertise and interest to support them in their quest for answers. The knowledge representation can guide system designers in determining what kind of information to provide and how best to do it, and can also serve as a resource for the users themselves, showing them how the “statistical world” is put together.

Key Words: Metadata, Ontology, End user services, Dissemination, Disintermediation

1. INTRODUCTION

A knowledge representation is a collection of symbols and rules for manipulating them that serves as a surrogate for objects and processes in the real world. Issues in knowledge representation can be traced back many centuries, including fundamental questions such as, “what will the representation be used for?”, “what should be represented?”, and “what should the representation look like?” Sowa (2000) defines knowledge engineering as “the application of logic and ontology to the task of building computable models of some domain for some purpose” (p. 132). It is this aspect of knowledge representation that we address in this paper; specifically, the role of knowledge representation in supporting the use and dissemination of statistical information. There are many names for knowledge representation and its close relatives, including metadata and metadata registries, knowledge stores, enterprise modeling, and knowledge management. Regardless of the name, knowledge representation can serve as a powerful and flexible foundation for the provision of statistical information to end users.

This paper has its origin in work done for the United States Bureau of Labor Statistics (BLS) to answer the question, “Can a knowledge representation of BLS statistical information help support its use and dissemination?” We investigated the feasibility of developing an overarching knowledge representation for BLS information that captured its semantics, including concepts, terminology, actions, sources, and other metadata in a uniformly applicable way. This work was motivated by the need for a foundation of information on which a variety of tools to help end users of the agency’s information products, including its Web Pages, could be based. These tools could help users learn what information is available, find the information they need and develop a better understanding of what the information means. For example, the BLS Selective Access application allows end users to “formulate and submit a query to the BLS LABSTAT database for the purpose of extracting timeseries data” (http://www.bls.gov/sahome.html). Who is likely to use such a tool, and what kind of help should the BLS provide for them? What information resources will the help services draw upon?

The last few years have seen the development of tremendous interest in the application of knowledge representation. Several forces have combined to bring this about.

• The need for an organization to recognize what information resources it “owns” and how they are organized.
• A recognition of the benefits of data sharing both within the organization (i.e., among different departments) and among related organizations.
• The desire to leverage information resources for as many purposes as possible; repurposing existing resources.
• The need for data/information dissemination both within the organization and to external users. The Web creates a convenient, but disintermediated dissemination platform, that is, one with no person such as a reference librarian immediately available for consultation. Organizations need to “help end users help themselves”.

961
The idea of putting all an organization’s information into a single coherent knowledge representation is attractive, but is probably not entirely realistic. An organization must examine its own information resources, and determine what goals are most important. In this paper, we will focus on the representation of knowledge that can help support dissemination of information to end users. (Note that “end user” also includes those within the organization.) First, we’ll discuss some of the problems of providing access to statistical information for end users. Next, we’ll present some of the varieties of knowledge representations along with some criteria for evaluating their effectiveness for an organization. This will be followed by a discussion of the kinds of services the representation could support, drawing examples from the BLS. Finally we’ll present some issues and strategies involved in an organization's development and adoption of a knowledge representation.

2. PROBLEMS IN THE USE AND DISSEMINATION OF STATISTICAL INFORMATION

As organizations such as the BLS make more of their information products available to end users over the Web, they must also make more support services available. The tasks that users must perform themselves, rather than obtaining results as published end products, what the users actually know about their information needs, and the information available to them all suggest that organizations should examine the kinds of tools, services, and support they provide.

In theory, the user population of the BLS site is the same as the user population of the Web itself. In practice, this isn’t the case, but there is more opportunity to attract casual browsers or people with mild curiosity about the information rather than those with direct information needs. Before the Web, for example, a high school student looking for information about the inflation rate might have been content to find something in a (possibly out of date) reference book, rather than trying to find a government publication or actually calling the BLS for the most current information. Now, however, the same student may find looking for the information on the Web more fun.

Unfortunately, looking for information on the Web often puts a greater burden on end users. First, there is the problem of finding the appropriate Web site, and the appropriate page within the site. This generally requires framing the information need as a query, identifying one or more search terms that will match the terms used by the organization itself. The end users must also be able to recognize when the retrieved information actually answers their expressed needs, as well as whether that matches their underlying information need (Taylor, 1968). Without a helpful intermediary, such as someone on the answer line from the organization itself, the end users may feel quite on their own. Assuming the end users have found the appropriate information, additional tasks may appear. There may be additional manipulations or operations to tailor the data to their individual needs (as with the BLS Selective Access page mentioned above) or to format it for downloading or printing.

Finally, there is the problem of interpretation. This is especially crucial with statistical data, regardless of the user population. Non-expert users may not fully understand what the data represents, and may therefore draw unsupported conclusions. Expert users may require additional information, such as the specific make-up of the user population that provided the data, or any seasonal adjustments made, in order to make best use of it, or even to decide if it is usable for their purposes. This problem stems from some characteristics of statistical data. Statistical information is more than just a table of numbers; the units, the column and row titles, processes such as normalization, imputation, and aggregation, sources of error, sample population, and many other factors all contribute to the meaning of the data. Real understanding of the data depends on understanding all these supporting information. Of course, all users do not require deep understanding; the high school student may be best served with a brief definition of inflation and a tutorial-style paragraph describing how the Consumer Price Index (CPI) is calculated. The economist planning pricing structures for a company’s products may require a much more sophisticated understanding.

The problem for organizations that provide statistical information on their Web site to a variety of users can therefore be summarized by the following questions.

- What kinds of tools or services are needed to support end users’ search for information?
- What kinds of tools or services are needed to support end users’ manipulation of statistical data?
- What kinds of tools or services are needed to educate users and help them understand what they have found, at a level appropriate to their expertise?
These three questions lead naturally to the next question.

- What kind of knowledge about their own information do organizations need to support these tools and services?

The answer to this last question could include many kinds of knowledge structures.

- Dictionaries or glossaries of vocabulary specific to the organization’s domain.
- Thesauri, incorporating the structural relationships among vocabulary and/or concepts (e.g., synonym, broader term, narrower term).
- Crosswalks, mapping specialized domain terms to general vocabulary, such as that used by a non-expert end user or mapping between different uses of terms within the organization itself.
- Authority or ownership information. Who is responsible for maintaining data and information artifacts? Who has the authority to make or approve changes?
- Metadata, which could include the source of the information, the date it was collected, its format, etc. In the statistical world, this could include the instance of the survey from which the data comes, and perhaps a brief description of any imputation, etc. that it had undergone.
- Descriptions of common processes, such as formulae or algorithms used to produce data items, series, or other statistical products.

This is information that most organizations already have in some form or another, however, it may not be readily available for providing the kind of integrated support that is needed for helping end users, sharing within different areas of the organization, or sharing across organizations. Information may be in different software packages (spreadsheets, database management systems, text files), in different data structures, or even stored in experts’ minds, and not actually recorded anywhere. The goal of incorporating this information into a single, coherent, well-understood knowledge structure is to make it readily available in a flexible way that will inform a variety of tools and services, both computer- and human-based.

3. KNOWLEDGE REPRESENTATION FOR STATISTICAL INFORMATION

In this section, we highlight some characteristics of knowledge representations, along with some criteria by which they could be evaluated for appropriateness and “fit” with an organization’s goals.

Brachman (1979) proposed five levels of knowledge representation based on an analysis of a variety of representational structures that all went by the same name, semantic networks. Sowa (2000) found these levels useful in tracing the progression from the basic representation of natural language through computer programs that produce a useful product. The levels are also helpful in identifying the type of information that is useful for the applications described in this paper.

1. *Implementational.* The level of data structures such as atoms, pointers, lists, and other programming notions.
2. *Logical.* Symbolic logic with its propositions, predicates, variables, quantifiers, and Boolean operators.
3. *Epistemological.* A level for defining concept types with subtypes, inheritance, and structuring relations.
4. *Conceptual.* The level of semantic relations, linguistic roles, objects, and actions.

For the purposes that we are focusing on here, the most appropriate level is the conceptual level, perhaps augmented by information at the epistemological level. The ISO/IEC 11179 metadata registry standard supports this idea.

A conceptual data model describes how relevant information is structured in the natural. In other words, it is how the human mind is accustomed to thinking of the information. It is one layer more abstract than a logical data model that describes a particular computer-based system. The conceptual data model provides an excellent place to start modeling data within the sphere of interest. It is also the most viable level at which to integrate different data models because object representational differences are excluded. (ISO/IEC 11179, Foreword).
The metadata and knowledge management research community has drawn extensively from research in ontology. The scope and purpose of an ontology is generally extended when it moves from being a static representation of a set of concepts and their relationships, to being the living repository of an organization’s operational information. Gruber (1995) discusses the requirements for such ontologies. First, he states the importance of evaluating the quality of an ontology based on its intended purpose. “Formal ontologies are viewed as designed artifacts formulated for specific purposes and evaluated against objective design criteria.” (p. 907) One could debate the probability of developing truly “objective design criteria”, but it is clear that the designers should have a good idea of (1) the kinds of information that will be stored in it, (2) who the intended users are, and (3) what they will be doing with it. He suggests that design criteria should include clarity, coherence, extendibility, minimal encoding bias, and minimal ontological commitment. Especially with the goal of information sharing, these criteria all aim to promote flexibility of use. Gruber also mentions the importance of consistency within the ontology. In many situations, however, there may not actually be consistency in the use of concepts and terms within the organization. The BLS pay concept family was selected as the focal point for a preliminary investigation of BLS concepts and terminology in part because it was a complex concept of interest to a wide community of users and other agencies. More importantly, it was commonly acknowledged with the BLS that different programs used different definitions of the concept. For example, one categorization of pay concepts could be based on whether they were included in or excluded from definitions of income used in two surveys, the Current Population Survey (CPS) and the National Compensation Survey (NCS). Rather than a forced consistency, the goal may instead be consistency within each context of use, and clear identification of when each definition applies. Guareno (1997) suggests the importance of domain knowledge in ontological modeling; this may be especially true when the model is intended to be shared among organizations.

ISO/IEC 11179 Specification and Standardization of Data Elements is a standard that specifies the structure of metadata registries. The ANSI standard ANSI3.285, Metamodel of the Management of Sharable Data (which is to be included in a revision of Part 3 of 11179) is a closely related standard. According to the ISO/IEC standard, a data registry is … a place to keep facts about characteristics of data that are necessary to clearly describe, inventory, analyze, and classify data. A data registry supports data sharing with cross-system and cross-organization descriptions of common data. Units of sharable data have precise identifiers, meanings, structures, and values. They are consistently deployed among users and systems and are centrally administered within an organization. Data registries may be organized into federations for interchange among many enterprises. A data registry assists users of shared data to have a common understanding of a unit of data’s meaning, representation, and identification. Just as a data registry may assist users in understanding like units of data, registries also assist in the understanding of differences of similar but different units of data. (ISO/IEC 11179, Section 1.1)

A metadata registry is thus a data registry designed to hold the metadata that describes an organization’s data. Note that the term “knowledge representation” or “knowledge structure” is more general, and could refer to the organization’s data or metadata. (Of course, one person’s data is another person’s metadata.)

Jurisica et al. (1999) propose four categories of ontological information that could be of importance in modeling information systems.

**Static ontology** describes things that exist, their attributes and relationships. **Dynamic ontology** describes the world in terms of states, state transitions, and processes. **Intentional ontology** encompasses the world of agents, things agents believe in, want, prove or disprove, and argue about. **Social ontology** covers social settings, permanent organizational structures of shifting networks of alliances and interdependencies. (p. 485).

An examination of these categories could help an organization decide where to put their efforts in developing a knowledge management system, based on the type of information and related functions on which they wish to focus.

In the statistical world, the static ontology might encompass concepts such as surveys and survey instances, questions, data items, aggregates and indexes, and the terminology and meanings that are associated with them. The dynamic ontology would represent processes concerned with administration of surveys and other types of data collection, data manipulation functions and the actions required for publication and dissemination of results. An understanding of the processes which produced the statistical data is vital to complete understanding of its meaning. The concept of time would also fit in here, central to the production and understanding of time series, among other products. The social ontology can also play a role in supporting the use of statistical information in at least two ways. As organizations seek to share information, either directly or through mappings such as that described in Kim
et al. (1999), the networks of communication among organizations could be represented here. Another important part of representing an organization’s information is determining and recording which departments or people have responsibility for maintaining it, including authorizing changes or updates to it. In the ISO 11179-3 metadata model, this notion is represented by the “administered component” entity, which connects data items of all kinds to an administrative entity. The role of the intentional ontology is less clear, especially in supporting user services. Juricica et al. (1999) state that it “allows alternate realities to be expressed and reasoned about” (p. 487). This would occur, for example, when a researcher postulates competing hypotheses or explanations for a result. How crucial modeling this aspect of information use within the organization’s information system, however, is debatable. Based on this overview, the static and dynamic ontologies would be the most important aspects of knowledge representation for supporting access to and use of statistical information. (Note the similarity to Brachman’s conceptual and epistemological layers.)

Criteria for evaluating the appropriateness of a knowledge organization should be based on the characteristics of the organization’s information as well as the uses for which the representation is intended. Criteria proposed for the BLS knowledge organization grew from examination of these characteristics as well as research in metadata structures, conceptual structures and ontologies, and other types of knowledge representations.

1. The representation should be compatible with other organizations’ knowledge structures. This is vital for supporting data sharing among organizations, an important consideration for the BLS. It already shares information with other government agencies such as the Bureau of the Census, and more data sharing is envisioned in the future. This criterion suggests that a “home-grown” structure is not ideal – using an existing structure, especially one conforming to an international standard (such as ISO/IEC 11179) is preferable. Further, in adopting any standard or model for a technology that is intended to promote the sharing of data (e.g., EDI, network standards, etc.) there is a critical mass of adopters that must be reached to achieve real success. (The Web page for the current ISO/IEC 11179 draft includes a good discussion about the role of data models in data sharing.)

2. The representation should be capable of representing not just core concepts and definitions, but also the relationships between concepts; modification, source, aggregation, etc. This is another facet of the central theme that statistical information is more than just a number.

3. It should be able to associate concepts, such as wage with the operational data, (i.e., that gathered through surveys), that are their instantiations in the real world, such as dollars/hour.

4. It should be able to associate information about processes and functions that apply to data elements, such as those used in imputation or aggregation.

5. It should support a wide variety of information processing functions and procedures used both within the organization, and by it to make information available to the public. Examples of functions within the agency might include recording changes in definitions, formulae, or survey questions, creating a glossary, or determining which program was responsible for maintaining a particular data set. Examples of functions which make information available include publishing data along with their definitions, limitations, or other associated information, developing tools to aid end user access to agency data and supporting information, and providing various help and education facilities, so that end users can learn more about the information the agency has.

6. It should include structures recording terminology and thesaurus relationships among terms. Terms that are ambiguous within the organization, or between technical and nontechnical uses should be clearly identified as such, and the contexts in which the different definitions (and all the other information connected with the usage) apply should be described. This information also clearly separates the “authority” terms and definitions sanctioned by the agency from the end user terms and definitions.

7. Thesaural relationships should include the familiar ones such as synonym, broader term, or narrower term, that are useful with precisely defined terms and term hierarchies. But they should also include those that are useful in building “crosswalks” between end user terms and agency terms. These relationships are likely to be somewhat “fuzzier”, as the end user definitions themselves are often defined in a more general way. Such relationships might include overlaps, near synonym, or often confused with. The structure that these
relationships produce is frequently not the familiar one-to-many structure that results from the broader term – narrower term relationship, but rather a many-to-many mapping that creates a complex network. The crosswalk structure lets end users start with whatever terms they know, and then provides guidance to steer them to the agency terms, and thus, the information they are seeking.

8. It should require as little effort as possible initially to populate it with data. The development and population should be based on the requirements analysis for the needs of the organization, with priority given to what is needed to support the most important tools. In the case of supporting end user access to statistical information, for example, concepts, terms, and definitions might be most important. For supporting expert use of statistical data, on the other hand, metadata concerning the provenance of the data might be more helpful.

9. The representation should make updating existing information as simple as possible. In some cases, updates may be automatic (or nearly so), such as adding metadata for new editions of surveys. Other kinds of updates, such as redefining a term or concept to reflect new usage (e.g., enlarging the geographical area covered by a metropolitan region) may require human effort. There should be a minimal duplication of existing information. To this end, the representation should not necessarily require that all information be maintained directly in its structure. Rather, the contents of the representation could be pointers to resources that exist within the organization.

A knowledge representation that satisfies all these criteria brings with it powerful advantages for the organization. However, it is important not to overlook the costs that also come along; the need for maintenance, the potential difficulty of incorporating it into the organization’s processes, and the need to get buy-in and support from key decision makers.

4. SUPPORTING USE AND DISSEMINATION WITH KNOWLEDGE REPRESENTATION

In Section 2, we proposed three types of tools or services that an organization might need to provide in order to better support end user access, especially in disintermediated modes such as Web pages.

- Tools or services to support end users’ search for information
- Tools or services to support end users’ manipulation of statistical data
- Tools or services to educate users and help them understand what they have found

In determining what tools and services to provide, it is important to consider users’ needs, domain expertise, and familiarity with the data. An external expert who works intensively with the information may have a higher level of understanding than an internal employee who is only peripherally involved with the production and dissemination processes. In this section, we present some examples of how knowledge representations of various kinds can be used as the basis for these tools and services.

Soergel (1999) claims that, “Ontological and lexical structures are the underpinning of scientific and scholarly work, of learning, and of machine intelligence. They serve many critical functions in thinking and in communicating, organizing, and retrieving information by people and machines.” (p. 1119) Among the specific activities they can support, he includes many that are crucial to the dissemination of information such as helping users understand the structure of the domain, concepts used in the domain and the terminology used to refer to them, and supporting search and access to information.

Kim et al. (1999) discuss the search problem of trying to match “ordinary language” with domain-specific controlled terms, such as that specified in a metadata vocabulary. An example in their paper draws from U. S. Foreign Trade Import-Export data. An end user wanting to find data on “cars” (in the normal automobile sense) who entered that term would instead end up with information on “Railway or Tramway Stock”. The appropriate data set can be found under the term “Passenger Motor Vehicles, Spark Ignition Engine”. One source of this problem, especially with statistical databases, is that the vocabulary allows for very few access points in comparison with bibliographic databases, where one may search title, abstract, and subject keywords. Data items often do not have synonyms or alternate authoritative names, and there may be little additional textual description. Kim et al. propose a solution that utilizes existing terminologies and metadata, and establishing mappings between them. In essence, an end user can start with a general language query, which is mapped to the Standard Industrial Classification (SIC), which is then mapped to the Harmonized System (HS) for classifying goods in international trade. The existing knowledge
representations, even though they were produced by different organizations for different purposes, can still work together to provide crosswalks between end user search terms and the “official” names for data items and sets.

Next, we present three examples of how a knowledge representation can support the use and dissemination of statistical information. The first focuses on terminology, the second on definitions, and the third on the information production and how the process may change over time.

Mapping Terminology
It has long been recognized that the terminology used by experts in a domain, for example, statisticians and other experts who regularly produce or use BLS data, is not necessarily the same as that used by non-experts, average citizens who use the BLS Web page to find information. This is similar to the problem addressed by Kim et al. (1999). The importance of this issue for easing access to information resources must be emphasized – if someone does not know the “correct name” of something, it can be very difficult, if not impossible, to find. As more and more agencies and other organizations are providing Web access to their data, they are realizing how crucial it is for them to provide some kind of terminological support for the end users; a “crosswalk” between the names and terms that are used inside the organization, and those that are used by end users trying to find information. Frequently, a thesaurus, glossary, or controlled vocabulary already exists in some form within the agency. Although this “experts’ thesaurus” may be helpful to users in finding good search terms, it does require some level of knowledge even to recognize an appropriate term. The crosswalk builds upon the thesaurus by linking terms that users commonly use for searching to the “official” terms. For example, most citizens are quite familiar with the term “inflation”; what they may not know is that the data item they are probably looking for is known as the “consumer price index”. Terms used by end users may be gathered from the logs of Web page search engines, news publications or broadcasts where they are used, and interviews with intermediaries and consultants who traditionally provide reference services. The end user terms can then be clustered into synonym and near-synonym groups, and the groups mapped to the appropriate agency terms. We will be creating a crosswalk of this nature for the BLS Selective Access page in the coming year.

Providing Definitions
An extension of the preceding example is to provide not only lists of agency terms for users to browse, but also the terms’ definitions. If the terminology has been organized into a structured thesaurus, then the definitions of the individual terms can be augmented by structural relationships, such as “broader term”, “narrower term”, “synonym”, etc. This type of information can help users in their search for information in at least two ways. First, they can find out what terms mean and how they relate to each other. Second, it helps end users learn more about what kind of information is available (and what isn’t), and how the “statistical world” is structured.

Modifications to a Statistical Concept
A user wants to find out about the unemployment rate. He or she could be interested in the figure for the most recent month, a time series for the past year, a time series for a particular month over many years, and so on. Both the particular information need of the user, and his or her knowledge and expertise in labor statistics affect what the “best” answer is, that is, the one that satisfies his or her information need.

- A high school student needs to find out the most recent employment rate for an "in the news" oral report for a Social Studies class. The need could be satisfied by the latest news release containing the sentence, “The number of unemployed persons, 5.8 million, and the unemployment rate, 4.1 percent, were essentially unchanged in October.” This rather isolated piece of information may be sufficient.

- A student in a beginning economics class needs to find the monthly unemployment figures for the past year. At this point, the concept of “seasonal adjustment” enters the picture. The student may need to know that unemployment figures are adjusted for the purpose of eliminating predictable variations in the employment figures such as major holidays or harvest. The user needs to know of the existence of the seasonal adjustment, and should also have some understanding that it makes it easier to observe nonseasonal movements in the monthly figures (BLS Handbook of Methods, http://www.bls.gov/opub/hom/homch1_h.htm).

- A labor analyst requires the unemployment figures for the last two decades in order to forecast trends for the next decade. Not only must this user understand what the seasonal adjustment is and how it is calculated, he or she should also be aware that its calculation has changed over time, and be able to determine the effect of that change on a long-term time series.
These brief scenarios illustrate the range of supporting information that an agency such as the BLS must provide the consumers of its information along with the “bare statistic”. The high school student may need none at all (at least for his or her immediate purpose), while the expert labor analyst may require details about how the figure is collected, calculated and adjusted, and how the procedure has changed over time. The knowledge representation in which this information is stored plays two important roles here. First, it links a variety of information together; the name of the statistic (unemployment rate), its basic definition, the time period and range for which it is collected (monthly, for the past six decades), the formula for calculating it (including seasonal adjustment), and the change history for the definitions and procedures. Second, the types of links and relationships among these various types of information tell the end users what supporting information is available, and help them decide if it might be useful. We call this the principle of successive revelation. End users should be informed of what information is available, but not be forced to see information they’re not interested in or not really capable of understanding. The relationship links among the various facets, e.g., “collection frequency”, “definition”, “formulae”, “adjustment formulae”, and “change history” could be used as labels on the Web page to inform users that more information is available, and how it relates to the basic statistic.

The information required to support all of these end user tasks exists in some form or another at the BLS and indeed, in most organizations. It may be found in handbooks, publications intended for the public, footnotes to column headings, or instructions for administering a survey. Perhaps the worst case is if it exists as “common knowledge” shared by the experts. An organization must decide whether the effort of pulling it together into a coherent knowledge representation is justified by the benefits of providing better information for its end users.

5. DEVELOPING A KNOWLEDGE REPRESENTATION

The process of designing and adopting a knowledge representation requires attention from various parties in the organization. Success requires coordinated effort along several dimensions, including developing a solid understanding of what the contents of the representation will be and what services it will support. It is crucial to get buy-in from pivotal people and departments. A use case can illustrate the role that the representation can play in important processes, such as providing tools or services for end users. The scenarios presented in the previous section could serve as the starting points for developing full use cases.

A use case is a narrative, like a story, that describes what an information system should do to perform one of its functions, or to achieve a goal. More formally, “a use case is a description of a set of sequences of actions, including variants, that a system performs to yield an observable result of value to an actor” (Booch, et al., 1999, p. 222.). The collection of use cases for a system describes its repertoire of functions – what should happen in order to get things done. Use cases are a component of UML, the Unified Modeling Language, which is rapidly being adopted in the information industry as a way of describing information system specifications. (See http://www.rational.com/uml/index.jtmpl?borschtid=08223005192316484400 for further information on the UML. Booch et al., 1999 and Alhir, 1998 are also good sources.)

Use cases can describe sequences of actions at any level of generality. A coordinated set of use cases can describe a function at a high level, and then decompose it into more detail. For example, one could have a general use case that describes the sequence of actions for accepting a user question and providing an answer. More specific use cases could then describe different kinds of questions, e.g., requests for data sets, requests for publications, requests for explanation or clarification of a term or concept, etc.

A use case describes what should happen, not how it should happen. A use case describes these aspects of the function.

- **Who the actors are.** An actor may be a person or a component of the system, such as a database or a metadata registry.
- **What the actors have or know as they enter the process.** For example, a user looking for a specific piece of information may know the appropriate agency term, or may know a common name for it.
- **The expected sequence of events that should occur under normal conditions.**
- **Alternative sequences of events, such as what happens if an error occurs.** For example, what should happen if the user enters a term that has no match in the database?
- **What the results of the process should be.** In the search example, the result could include a specific piece of information, or a pointer to the location where that information could be found (e.g., a URL). Results could also include an opportunity for the user to modify the query and try again, or to obtain online help.
The basic form of the use case is very approachable. Initial development depends more on familiarity with the functions and actors, and less with the details of how to make it work or exactly what the format of the data types should be. Use cases provide a good basis for “walking through” a procedure, determining what the possible actions are, what resources are needed, identifying what system components already exist, and what components need to be developed. Alternative use cases can be developed for the same function, to promote discussion of their comparative advantages. Because of these characteristics, the use case is a good tool for informing people what the role of the knowledge representation would be, how it relates to the organization’s other information resources, and how it could benefit the organization. Use cases also provide a foundation for developing new uses for the representation.

Sowa (2000) points out that, “Different applications require different amounts of detail in the ontologies and different levels of logic in the representation” (p. 412). If one knowledge representation is to support a variety of services or applications, it must be detailed enough for the most demanding, yet allow a more general level of treatment as well. It is also important to recognize that in most organizations, knowledge resources have been built up, more or less systematically, often over a period of many years. Different areas may have been developed by different departments, programs, and for different purposes. There may be overlaps in coverage, and there may be actual contradictions in meanings and use, even within a single organization. (Recall the BLS pay concept.) As Jurisca et al. (1999) state, “In each domain, there are points at which formalization becomes more of a straitjacket than a liberating force. The challenge is therefore not so much to decide which approach is better, but to develop techniques for the various approaches to work closely together in a seamless way.” (p. 491). Although there are standards or “customs” for various types of knowledge representations, ultimately each organization must decide what will best serve its own needs, taking the following issues (among others) into consideration.

- What kinds of user services (inside or outside the organization) need to be supported? Are some of higher priority than others? (The entire representation doesn't have to be built at once.)
- What kinds of knowledge resources already exist within the organization that could be incorporated into an enterprise-wide structure?
- What is the nature of the organization’s data and metadata? Does it cover a lot of conceptual ground, or is it more narrowly focused? Is it consistent throughout, or are there contradictions?
- Are there other organizations with which data or metadata will be shared? If so, in what state are their information resources? Can the development and maintenance effort also be shared?

6. CONCLUSION

This article has given a brief introduction to the topic of knowledge representation. We have focused on how a knowledge representation can provide supporting information for end users of statistical information, thereby improving its dissemination and use. In doing so, we have emphasized the following ideas.

- Statistical information is more than just the numbers; users must understand its metadata, definitions, derivations, and related information. The knowledge representation provides a structure through which all this information can be linked together.
- End users have varying needs for this supporting information, based in part on their individual information needs and their expertise with the statistical data.
- The organization should provide its users with the supporting information in order to further their understanding of the statistical data.
- Users should be informed that supporting information is available, and how it relates to the data, but should be able to decide if they want to see it, and if so, how much. We referred to this as selective revelation.

A knowledge representation is not a magic talisman; by itself it will not make the use and dissemination of statistical information easy and foolproof. But it can serve as an important component of tools and services that can help the users of statistical information find what they need and understand what they find.
7. REFERENCES


ABSTRACT

Government statistical agencies and other organizations publish data on the World Wide Web, making it available to users through browser-based tools. As a result, many data users with little experience working with statistical information now have online access to these data. Findings of usability tests of three Bureau of Labor Statistics and Census Bureau data access tools revealed usability problems in three categories: insufficient guidance or instructions for tool use, unclear data naming and database organization, and unsystematic interface organization. Specific instances of each problem class are provided, and methods for resolving these types of usability problems are proposed.

Key Words: World Wide Web, Access to statistical data, Usability

1. INTRODUCTION

Statistical data collected and stored by many public and private organizations can now be accessed through the World Wide Web (WWW). In the case of official statistics, data users access statistical information at the national level (e.g. FedStats in the U.S., the Statistics Canada site in Canada and the National Statistics site in Great Britain), or at the individual agency level (e.g., the Bureau of Census or the Bureau of Labor Statistics (BLS) sites in the U.S.) More powerful websites to enable users to access statistics simultaneously from various organizations are beginning to appear. For example “The Dataweb” now lets users access data simultaneously from databases at the Bureau of the Census and the National Center for Health Statistics and other agencies. University-based organizations such as the Inter-university Consortium for Political and Social Research (ICPSR) maintain on-line data repositories. Finally, sites operated by commercial firms offer other gateways to statistics on the Web. (URLs of some of these websites appear in the Appendix.)

Any of these sites offers users access to an enormous amount of data. The data are diverse in their conceptual bases, their content, their quality, timeliness, and many other attributes. Web access opens up all of these complex data resources to many potential users who may know very little about data at all, and are unprepared to cope with the volume and complexity of the information suddenly accessible to them without substantial on-line support and guidance. Several U.S. federal agencies have created software “tools” that aim to provide such on-line support to persons trying to locate and use data via the Web. The centralized FedStats website identifies nine such agencies, and some of these provide several tools.

The authors’ informal examination of the tools on this list and identified numerous usability problems, some of which appeared to pose serious barriers for users. On the bright side, several federal agencies, including the Bureau of Labor Statistics, the Census Bureau, the Energy Information Agency and the National Cancer Institute are currently performing usability engineering activities in order to make the tools at their websites more usable and useful (e.g. Levi and Conrad 1997). Nielsen (1993) characterizes usable software as easy to learn, efficient to use, easy to remember, promotes graceful recovery from errors, and is satisfying to use. This paper focuses on some of this usability activity.

1.1 Scope of this Paper

This paper reviews usability tests that the authors and their colleagues at BLS and the Census Bureau have performed to evaluate the extent to which some of these agencies’ data access tools are usable. The tools are displayed in a web browser and users interact with them in order to obtain agency statistics. The goal of the

---

1 We thank Cathryn Dippo, Carol Hert and Michael Levi for valuable comments on an earlier draft of this paper. The opinions expressed here are those of the authors and not of the Bureau of Labor Statistics.
research, of course, was not simply to uncover usability problems but to serve as a basis for improving tool usability through design modifications.

The tools tested were:

1. FERRETT (“Federal Electronic Research Review Extraction & Tabulation Tool”)—This tool is hosted by the Census Bureau but enables access to survey data at Census and other agencies. It has some capability for online data processing, e.g. tabulation as well as a download function.

2. A Java data query tool—This tool helps users access any of the BLS “time series,” i.e. data points for a given sub-population collected periodically (monthly, quarterly, etc.) over a period of time. A prototype version of this tool was tested on BLS’ collection of labor force time series. This is the agency’s largest database of this type, and contains over 50,000 such series. The labor force database was chosen for test purposes because of its size, on the assumption that a tool shown to be usable with such a large database will be robustly usable when it is scaled down to give access to smaller databases.

3. A “wage finder” tool—This tool is in many ways a variant of the Java data query tool that applies to many different datasets. The “wage finder” gives users access to wage data (point estimates of hourly wages) from BLS’ National Compensation Survey. Metropolitan area and occupational types as well as specific occupations are used to categorize these data.

Usability tests on FERRETT began after it had been deployed for some time as a CGI application, and continued with a new Java-based version whose design grew out of the usability test program. This paper draws on test results for the original CGI version. Both of the BLS tools underwent usability testing and design improvement based on the tests while still in prototype form and before public release.

The functionality of the various current access tools varies considerably. While some of these differences undoubtedly reflect the influence of real differences in the nature of various organizations’ data, the authors propose that, from the user’s perspective, the task of specifying and accessing data is, essentially, the same irrespective of such variations among different data sources. This forms the basis for conceptualizing problem types across different tools. If there were some convergence in developers’ conception of the users’ task, as there is for applications like spreadsheet or word processors, users could learn to use data access tools more easily and transfer their data access skills from one tool to the next.

### 1.2 Conceptual Framework: A Generalized Data Access Task

This sub-section describes the authors’ conception of a generalized data access task. (A relevant workshop report can be found in Levi and Conrad 1999.) Users are essentially narrowing down the available data to obtain a subset—often just one data point or series—of interest. Users must perform three component tasks sequentially to reach this goal. First, they must specify the attributes of the data in which they are interested to the extent that the data access tool allows, and send this description to the appropriate database. Any usable tool should transparently formulate a query from the user’s specifications that the “back end” database software can interpret. The tool then displays for the user a set of descriptions (labels, names, etc.) of data elements or datasets that match the user’s data specifications.

Second, the user must evaluate the candidate descriptions (e.g. data series names) on the basis of how well they fit the given specifications, and choose the best match(es) relative to the task goal(s). The user may decide that none of the returned descriptions are acceptable but, if at least one is chosen, the tool lets the user request the chosen item(s), which are returned. If the user does not choose any she can either return to the first step and reformulate her query or abandon the search. Third, the user examines the actual data to determine if they fulfill her goal(s). If unsatisfied with the results, the user must then decide whether to return to the first step and modify specifications in hope of obtaining better results or abandon the search. The tests reported here included little attention to this third step, but future research may focus on evaluating usability from this standpoint.
2. USABILITY TEST METHODS AND PROCEDURES

This paper is based on three usability tests whose findings are summarized in the next section, along with recommendations about how to avoid the kinds of problems identified in the tests. Two of the tests were conducted at a desktop workstation in the BLS Usability Test facility. This facility places the user in a typical office environment with bookshelves, and a modular desk and table. A personal computer offers reasonably high-speed Internet access. The computer enables web browsing with a late version of either Netscape or Internet Explorer. Normally, the user is given an orientation and task instructions, then left to work alone while test personnel observe user actions on a TV monitor in an adjoining room. Cameras and a scan converter in the work room deliver images of the workstation screen, the user’s face, and keyboard and mouse actions to the monitor. It is possible to provide additional guidance, ask questions about apparent problems, and so forth via an audio intercom.

The test of the “wage finder” prototype tool was conducted on laptop computers at a hotel in a large southern city. This test took the software to the field to capitalize on access to a user group with high motivation to work with the NCS survey data. These were compensation specialists from both private and public businesses and labor unions who were attending a state convention. Test personnel sat beside participating users at a worktable, observing and guiding the users’ performance. No videotaping was possible under these conditions. BLS staff wrote down notes about their observations as the users worked on prescribed access tasks.

These tests have all involved small groups of from 5 to 10 users, although the FERRETT tool was tested on more than one occasion so that the total number of user participants grew to about twenty for that tool. Except for “wage finder” test participants were a diverse group recruited from academia, government, the press, and private business. Most volunteers have been at least moderately familiar with using statistical information, but the test team has attempted to include some users in each group who are unfamiliar with procedures for accessing data on the Web. Users were assigned a range of tasks to perform in the various tests. In most cases, specific data targets have been assigned to users so that the test team can measure users’ success. In a few cases users were invited to use the tool to look for data that were personally interesting to them.

We report observational and qualitative results, although in most of these studies we quantify the users’ performance, for example by counting how often they access the intended data. In some cases, test sessions were videotaped for later review, but often test data were captured primarily as notes taken by observers during the test sessions coupled with user feedback in debriefing interviews following the individual sessions. Test observers met following one or more sessions and literally “compared notes,” discussing their observations of signs of difficulty in using the tool until they could agree on a characterization of the problem and a formulation of its source in tool design. In some cases, these qualitative summaries were simply delivered to the software developers, who then came up with design changes aimed at improving usability. In other cases, usability experts on the team worked along with the developers on making design improvements for the interface. In either case, of course, the developers wrote the code to implement design changes.

3. SUMMARY OF USABILITY TEST FINDINGS

The authors reviewed the summary reports on all three usability tests and abstracted from them a small number of “cross-cutting” findings for inclusion and discussion here. There were numerous relatively minor usability problems that were specific to a given tool, but these were left out of consideration in order to highlight three major usability problem types that were plainly evident in all of the tests. These cross-cutting problem types are:

1. Insufficient guidance or instructions for tool use
2. Unclear data naming and database organization and
3. Lack of organization and feature prioritization.

A brief sub-section discusses each of these types of problems and offers suggestions for overcoming each type. The authors wish to emphasize that the tool designers have welcomed the insights the tests gave into problem areas, and have used the test findings in every case to rework the tool so that usability is now very significantly better for all of them. Sometimes the improvement was confirmed by a repeat test, but often improvement is shown by inspection.
3.1 Problems of Insufficient Guidance or Instructions for Tool Use

Given that the applications tested are characterized as “tools,” it was surprising and disappointing to the test team to find how little guidance the interface gave users on how to apply the tool to the data access task. This problem was particularly evident at the level of overall orienting guidance. Often the interface had clearly recognizable “widgets” such as drop-down lists, radio buttons, etc. and the specification alternatives these represented were shown clearly. The tests showed that inadequate description of the general context within which the tool is embedded was a major source of this problem type. Answers to questions like “What sorts of data can be accessed using this tool?” “How do I describe the data I’m looking for?” or “Where do I start on this interface in order to use the tool most effectively?” were often lacking. Sometimes important instructions were available in the “help” function, but de-emphasized. The link to some useful instructions for using FERRET (Figure 1) is labeled “Hint.” By embedding this link in a terse, inadequate instruction, the design implies that the brief instruction should be sufficient. In fact, the “Hint” clearly offers the user better, clearer guidance.

![Figure 1. Example—Problems of Inadequate Tool Use Guidance](image)

This type of problem can be alleviated if tool developers adhere to guidelines such as:

- Don’t assume that all users who find the site know much about the statistics available there. Provide high-level orienting information about the data the tool is linked to, especially contextual information such as who collects these data and for what purposes. Tell them the topical domain of the data – if the data are about mortality, for example, then say so. Don’t make users guess.

- Provide adequate specific instructions as well. Often tools have a surprising amount of unoccupied “screen real estate” available for more extensive guidance. On the other hand, something as simple as numbering the steps in a procedural sequence with “1,” “2,” “3” and so on takes little room on the screen—but can be highly effective. One of the versions of FERRETT used this simple, explicit “hand holding” and it improved usability, although eventually this feature became unnecessary after additional tool redesign.

- Well designed “help” screens can greatly enhance a tool’s usability. They should not however substitute for putting as much guidance on the primary interface as possible. Also, “help” should be provided in small units closely tied to specific contexts, not lengthy and irrelevant discourses full of general information. The tools included substantial “help” that users could access, especially the two BLS tools, but perhaps did not indicate clearly enough that help was available and how to access it.

- Some tools attempt to compensate for limited usability by offering extensive tutorials to help novice users acquire necessary expertise. Unfortunately users who need the tutorial are unlikely to spend the time to work through it, preferring instead to “jump in” right away in order to obtain real results This is especially true for web users who expect results in a small number of clicks, and are very unlikely to follow extended performance-based practice exercises or to engage in word-by-word reading of large chunks of text, according to Nielsen (2000).
3.2 Unclear Data Naming and Database Organization

This problem area also has two aspects, as the title indicates. One aspect involves unfamiliar terms from specialized vocabularies that data experts or computer experts embed in the tool. The other aspect involves failure to communicate to users in their terms how data names or descriptions are organized in databases. As an example of the first problem aspect, the data series that are accessible using the Java data query tool often include unfamiliar, specialized terms in the names of the series. These terms often make distinctions between data series that are meaningful only to expert users. A prime example in the labor force data is a distinction between two measures based on the same concept, e.g. “unemployment,” where one is expressed as a rate and the other as a level. To the non-expert user, the most meaningful part of the name in either case is the familiar term “unemployment.” The technical rate-level distinction may not be noticed at all.

Some distinctions may not only be difficult for non-experts to understand, but they may represent entities that are of little interest or value to broad segments of the total user community. An example of this would be age or income data series which involve unequal intervals instead of more familiar 5- or 10-year age groups, or income categories other than those using increments of, say, 5 or 10 thousand dollars to define their boundaries. The unusual categories may have some utility for a few specialist users, but for most users their inclusion in the set simply complicates the selection process by inflating list length.

The other aspect of this problem type, data organization, stems from the fact that the databases to which access tools apply are often organized and hence displayed in a manner that makes it difficult for users to process information (metadata) about database content. Some examples of how data organization can lead to designs that are hard to use are noted in Figure 2, a screen from BLS’ Java data query tool. This tool supports winnowing long lists of time series, over 50,000 in the case of labor force statistics, to find the series the user is interested in. The user does this by selecting various values (e.g. the seasonally adjusted number of men employed in agriculture) in text boxes, each containing either a list of statistics which differentiate among series as classes, or all possible values of one of the variables that series may include.

![Figure 2. Example—Problems of Data Naming and Organization](image)

“Measure” is the data type that defines a series subset such as unemployment rate, but it is not distinguished from series variables (e.g. Age). All are called “Characteristic.”

Repeating the measure label “Average Weeks Unemployed” for each series uses nearly half of each line, forcing users to scroll to view other series attributes.
Each labor force data series is identified in part by the name of a labor force measure, e.g. unemployment rate. Each measure identifies a broad statistical class of series. This name appears first in the list of defining descriptors for every series involving the measure, but it is not distinguished in any way from the additional descriptors that specify attributes or values of variables used in the series. This has two consequences for the tool designer. First, it requires the tool to indicate clearly to users that they must choose the correct statistical measure—that this is the crucial selection criterion. The design shown in Figure 2 failed to provide such an indication clearly enough. This led some users to pay more attention to specifying choices of the variable components like age and sex than to picking the basic measure defining the series type they were looking for.

Second, repeating the measure in every label for all its class of series can use quite a bit of screen area. The fact that this redundant descriptive term falls at the beginning of each series label makes it more difficult for users to scan the list of series names, since they have to skip over the measure name to find the terms that really distinguish one series from another.

These series names may contain a large number of descriptive terms. From the standpoint of clear labeling, a data access tool should enable the user to view the full name of each series in the currently selected set so that she can determine whether the results the system is returning are satisfactory. In Figure 2, the tool tries to do this, and furthermore show the list of selected items on the same screen. As the figure also shows, however, the text box cannot be wide enough to show long series names because the recurring measure label plus an arbitrary alphanumeric identifier for a series causes parts of long labels to be hidden. Thus the user must scroll left and right to view all information for some series, and this diminishes usability.

Tool developers should adopt practices such as the following to reduce the seriousness of data naming and organization problems. It should be noted, however, that as long as the data to which a tool applies continue to be named and/or organized in ways opaque to users, there is a limit to how much design can compensate.

- A usable tool should provide easy access to definitions of technical, unfamiliar descriptive terms. Such terms might be hyperlinked to a glossary (with easy return to the working interface of the tool), or mouse-over definitions could be used.
- The tool should inform or indicate to the user if a specifying term, such as the measure for labor force series, is a unique identifier for a data type (major sub-classification category). This kind of term should be distinguished from secondary attributes that are used to refine selections within the major data type.
- A usable tool will clearly indicate when users try to specify for missing or never-collected data. A tool can do this by showing missing data types or values in gray or in a different color from valid data specifications.
- A usable tool should make it easy for users to review the results of their specifications, to verify that they are getting the kinds of data they want. The Java data query tool shown in Figure 2 attempted to do this but its success was limited because of how series labels are constructed.
- Including features in a tool that give users the means to combine specification terms, such as “and-ing,” may seem an attractive way to gain efficiency in data specification tasks. Research shows, however, that many users have difficulty constructing queries that require Boolean logic, as in SQL (e.g. Greene and Devlin 1990) beyond simple “and-ing,” (Sewell and Teitelbaum 1988). At one point, the FERRETT designers included a feature intended to support a kind of graphical construction of SQL-type specifications by pointing and clicking data names and operator buttons, instead of using text entry. This feature did not help users do any better and was dropped from the design. Even if logical operations are performed implicitly, e.g. by “and-ing” individual terms chosen from different lists, users need clear feedback about the result of the operation. More generally, the interface should inform users whether their choices are narrowing (“and-ing”) or broadening (“or-ing”) the set of data specified.

3.3 Problems of Screen Organization and Layout

One of the authors reviewed about 200 “help desk” messages to the FERRETT support staff in order to get an idea of what kinds of tasks to use in usability tests of this tool. He found that a majority of FERRETT users who contacted the help desk have simple data access goals. For example, most are looking for data involving one or a few variables, often specific to a geographic area such as a state or metropolitan area. They may wish to perform some basic manipulation of these data, such as selecting only one gender category, restricting age values to a certain range, and the like; rarely do they wish to do more complex data transformations. Similarly, the field usability test of
the “wage finder” tool indicated that a key set of potential users—wage analysts—usually want to access an average hourly wage for just one occupation and one geographic region at a time. These observations seem to indicate that usable data access tools will offer a set of “core” features to support simple, very common data access tasks and clearly identify these on the tool’s main interface. Even if a tool offers some on-line analytic support, as FERRETT does, that too should be simple. Experience with FERRETT shows that most users will be satisfied with simple and straightforward on-line data tabulations, rather than applying any more powerful techniques such as regression analysis. Those users who do complex analyses of the data are normally experts who will download the desired subset of data to the desktop for such analyses, because of the convenience and flexibility of working off-line.

Figure 3: Example—Problems of Screen Organization and “Clutter”

Another way in which screen layout can facilitate users’ task performance is by using default specifications on the interface. In many cases, our usability testing clearly indicates that one selection is so popular that setting it as a default choice in the tool eliminates the need to change the setting for most users. In general, it seems plausible for most selection options to set the default selection inclusively, e.g. for “Sex,” data for both men and women. In some cases, one of the tools provided no default value for a specification, e.g. in Figure 2 the list box for “Sex” is empty. In such cases users hesitate to manipulate that interface feature. They may feel constrained to explore the feature, e.g. the blank list box, just to discover what choices it offers, even if this serves no purpose in reaching their task goal.

Figure 3 shows a screen shot of the prototype “wage finder” tool that BLS tested with a set of known wage data users at a statewide convention of these professionals. The list box titled “Select a Level” (lower left portion of the screen) permits users to select a federal government occupational (GS) level as one of the specifications for finding an hourly wage. When users opened the tool this box had no default selection. In fact, most users were not interested in getting wage data by level, instead opting for “Overall Average Wage.” The latter is an obvious default choice. Note also that a control labeled “Don’t know what Level?” proves to be a means of accessing some kind of “help.” A label making this specific purpose clearer seems more appropriate. Finally, the test showed that these users had little or no interest in a feature to support unselecting choices from a cumulative list. Indeed, these users typically had no need for any facility to make multiple specifications before accessing data, which the tool supports with the lower-right text box and associated controls in Figure 3. Users said that on the job they would usually look for only one wage at a time.
Some design guidelines that can help developers avoid creating tools with this type of problem are:

- **Consider providing alternative versions of tools for different groups of users who differ in their capabilities and goals.** BLS’ “Most Requested Series” and “Selective Access” are to some extent examples of this. “Most Requested Series” is a kind of “front door” to “Selective Access.” It may satisfy the needs of many users, and also teaches the non-expert user by example about the kinds of data series that are comprehensively accessible through “Selective Access.” This method of user familiarization with data can be superior to a lengthy tutorial that merely simulates working with the data.

- **Most users will not have a clear mental picture of the totality of data that may be accessible using a given tool, especially if the tool allows access to many different data sets or series, each with many dimensions.** Where needed to help manage complexity, the tool should make the structure of accessible data explicit by more effective use of graphical design in the interface. Instead of simply showing an alphabetical listing of available data series, the designer should cluster related series by (labeled) topic. This is consistent with Donald Norman’s (1988, pp. 54-80) broad design principle that suggests capitalizing on “knowledge in the world,” i.e. the interface, not just knowledge in the mind of the user.

- **Tool design should fully support data specification if possible, i.e. offer users a chance to specify all data classes, variables, or variable values.** To do this may require multiple screens if the data can be specified on numerous dimensions.

- **Give users clear feedback on the interface when specified data are not available at the site.** The tool should inform users of “dead ends” as early in the specification process as possible.

### 4. DISCUSSION OF FINDINGS

The most telling effect of tool designers’ apparent lack of knowledge about the needs and skills of less-than-expert users was a pervasive assumption that all (or most) users know much more about the data than they actually do. The influence of this error is visible in all three of the cross-cutting problem types. If tool designers believe that users generally know a great deal about the data, both structure and content, then they will not be highly motivated to include much orienting information or procedural guidance in the a data access tool. The tool will require users to work with ill-formed names and unwieldy database structures if designers believe that, in general, these factors will not hinder user performance. And designers who believe that users are skilled in gaining access to data, and that they attach great importance to doing complicated operations with the data they get, will do little to logically organize interface features so that they highlight the key steps in simple data access tasks.

The assumption that users know more about the data than they actually do is particularly detrimental to usability because the current generation of data access software tools is usually retrofitted to legacy databases in which the data are organized to optimize storage efficiency, not accessibility. This organization may not reflect how users conceptualize interrelationships among different variables. The data are often described in specialized terms quite unfamiliar to non-experts. Users are expected to understand these terms, e.g. the difference between “employment rate” and “employment level”. Users may need to understand the subtle differences between statistics that sound very similar but are collected in different surveys by different methods, e.g. wages published by occupation from the Occupational Employment Statistics (OES) program and from the National Compensation Survey (NCS). For example, OES data are reported at the national level only, while NCS reports down to the level of metropolitan areas within states.

Finally, legacy databases were until recently used only by experts. Non-experts got processed data products, such as tables and reports. So data access tools often still show a bias toward serving “power users,” while features supporting simple and rapid access to a few data points or series are relatively de-emphasized in the design.

Designing for all potential users may well be impossible. Nonetheless, if this is the goal – as one could reasonably argue should be the case for government statistical agencies who serve the general public– it may be useful to describe some broad strategies that could lead to greater general usability if applied consistently.

One approach to tool design advocated by BLS consultants who have been studying web data dissemination from a combined perspective of information science and computer-human interaction recommends creating multiple alternative interfaces for different user classes (Hert and Marchionini, 1997; Marchionini, 1998). This design approach in turn requires first identifying a set of user classes, say those classes who are most likely to want or need
the data in question, and then implementing different interface designs for each such group. An alternative approach to designing for all users is to design a single interface geared to users with the lowest levels of knowledge and skill. This will frustrate users with higher levels of knowledge and skill but will not leave anyone behind.

A variant of this approach is to design multiple data access tools for multiple well-defined sub-populations of users, making the consequences of this decision explicit in the interface of each tool. A given user’s skill and knowledge levels can much more safely be assumed to match one of these tools. BLS has implemented this approach to some degree. For example, LABSTAT users who are seeking a specific, well-publicized number can use an extremely simple but inflexible “Economy at a Glance” tool to access a pre-formatted table containing current data on various major economic indicators produced by BLS, such as the current unemployment rate. The data are accessed with a single click—but are limited to those the designers (with the blessing of BLS subject-matter experts) chose to include. These data are also point estimates rather than time series.

Other users seeking a popular, often-cited data series can use the “Most Requested Series” tool. This tool is less restrictive, but also assumes some familiarity with BLS time series instead of single data points. It allows users to select from a list of frequently accessed BLS time series. Users can select desired date ranges and other output options prior to executing the query. This makes it relatively easy to obtain entire data series—it requires a few choices and clicks—but the user is limited to a small subset of all possible series and can distinguish among them only the basis of date ranges and output formats. Finally expert users looking for a less well-known, possibly complex data series can use the “Selective Access” tool, which empowers (and requires) users to formulate and execute more extensive specifications of the series being sought.

The Census Bureau also takes a multi-level, multi-tool approach. That agency also offers a range of tools, from simple to more flexible and complex. Like those in LABSTAT, these tools provide data from the same data sources but vary in the flexibility they afford the user and the degree of user knowledge they assume.

In some cases, e.g. FERRETT, the Census Bureau’s tool for accessing Current Population Survey and other survey data, users are warned explicitly that the tool is for experts. In other cases they are not; for example, the “Data Analysis System” associated with criminal justice datasets stored at the University of Michigan claims that its on-line analysis process is easy enough for “first-time users,” when in fact it is requires extensive analytic skill. It is our impression that the kind of warnings about required expertise that are found with certain tools generally underestimate the degree of user knowledge of the domain assumed in the design of the tool.

5. CLOSING COMMENTS

All three types of cross-cutting problems described in this paper can be traced to a fundamentally flawed process for gathering requirements prior to initial tool design. In the course of these studies, the authors came to realize that the tools’ designers lacked the most basic information about the kinds of data access tasks many types of users wished to perform, and what features should therefore be embodied in their tool. Instead, expert data analysts employed as agency staff primarily guided the design of the tools. To be sure, some of this guidance was informed by staff perceptions of the needs of (usually) equally expert data users outside of government, such as academics, policy analysts and other elites. Little or no information was available about data use and needs of members of the general public, who gained the means to become data users once agency data were made available on the Web.

Web data access tools must be designed to meet the needs of the entire broad spectrum of data users on the Web, including first-time users. In order to accommodate this diversity, a process to collect these diverse requirements and communicate them to tool developers is badly needed. The diversity of web users is so great that it is probably impossible to get a complete profile of all possible data users’ needs. Establishing even more limited needs assessment process is a daunting challenge, demanding creativity to design and substantial resources to implement. But the research reported here indicates an urgent need for greater efforts to learn more about what features data access tools should offer to a greater variety of data users. There is every reason to do the best we can, even though we cannot do everything, in gathering more information about a greater variety of data users who visit our statistical websites.

The authors hope that this paper will persuade its readers in the statistical community (and beyond) that usability engineering activities, including usability testing, can help assure that future data access tools will be much more
usable and useful than the “first generation” now available on the web. In order to make future tools more usable, agencies urgently need to collect information about their entire user base’s priorities in the kinds of data they want, their familiarity with data in general and those data to which the tool will apply in particular, and other relevant information. Only then can the design process be grounded, as it should, in a clear definition of the characteristics of the group or groups of users for whom the tool is primarily intended. The difficult work of collecting information about user needs will support a much more effective process of designing usable data access tools.

Frequent usability testing as the tool evolves can keep the design relevant to user needs. Usability evaluation of a tool when the tool is ready for deployment can increase the chance that users successfully access the data they are seeking. A tool that is hard to use is more likely to lead users to look for data from alternative, less authoritative—but more usable—sources.

6. REFERENCES


Greene, S. L. and Devlin, S. J. “No Ifs, ANDs, or Ors: A Study of Database Querying,” International Journal of Man-Machine Studies, 32, 303-326


7. APPENDIX

Various websites afford access to statistical information. As of July, 2000, statistics in the United States can be found at FedStats at http://www.fedstats.gov. This is a portal site for over 100 federal agencies with statistics online. For Statistics Canada, the link is http://www.statcan.ca. For the National Statistics site in Great Britain the URL is http://www.statistics.gov.uk. At a sub-national level an individual agency may maintain its own website. For example, the US Bureau of Labor Statistics site is http://stats.bls.gov. An example of a university-based data repository is found at http://www.icpsr.umich.edu/index.html for The Inter-university Consortium for Political and Social Research at the University of Michigan. Finally many commercial organizations such as Berinstein Research at http://www.berinsteinresearch.com/stats.htm offer other gateways to statistics on the Web.