DCPAPERS

Statistical Metadata Needs during Integration Tasks

Sheila Denn, Stephanie W. Haas
School of Information and Library Science, University of North Carolina at Chapel Hill, USA
{denns, haas}@ils.unc.edu

Carol A. Hert
School of Information Studies, Syracuse University, USA
cahert@syr.edu

Abstract

As part of an NSF-funded effort to work toward a national statistical knowledge network, we have been working to determine what elements of statistical metadata are most crucial for end users as they attempt to find and use data from US Federal statistical websites. We report here preliminary findings from a study conducted to compare how experts and end users interact with statistical data and discuss the implications of these findings on the construction of architectures and interfaces designed to integrate statistical data across agencies.
Keywords: *descriptive metadata, government agencies*

1. Introduction

Any effort designed to integrate data for users from across different sources and domains necessarily depends upon the effective use of metadata to provide linkages between sources and to allow the user to orient herself within the data. In our efforts toward building a statistical knowledge network (SKN), we need to concentrate on those metadata elements that are most important to the end user in supporting the integration process. To this end, we constructed a study that was designed to reveal the kinds of issues and challenges that come up when users are completing integration tasks using statistical data, to identify the metadata elements associated with these issues and challenges, and to begin to construct a metadata architecture that prominently features the metadata elements identified.

2. Methodology

2.1. Goals and Research Questions

The goals of the metadata user study were to gain a better understanding of how experts and end users integrate information from statistical websites in order to complete integration tasks, and to explore how this understanding could contribute to developing tools and interfaces that would support users in their integrative activities. In particular, our research questions were:

- What problems/uncertainties do specific types of users have during tasks involving integration of statistical data?
- For the same tasks, what problems/uncertainties do experts perceive as being relevant to usage of the data by the user populations?
- How do problems experienced by end-users compare to those identified by experts?
- What metadata or other information can be identified to help resolve user problems?

2. Study Design

In this project, we used a scenario-based design approach to guide our work [1-2]. Scenario-based design is an iterative approach to system design that relies on user interaction scenarios, or narratives, as the source of guidance for design requirements. These narratives describe how an archetypal person (with a set of goals, behaviors, and knowledge) would carry out a series of interactions with a system. The articulation of the scenario enables designers to understand the features of the situation (e.g. needs analysis), determine appropriate system action (e.g. design requirements analysis), and document them [2]. The approach "exploits the complexity and dynamics of the design domain" [1] enabling designers to better understand real tasks and the constraints upon them. Because scenarios can be both concrete and easily changed, as well as written from multiple perspectives (e.g., from the viewpoint of multiple stakeholders) and levels of abstraction, they can provide guidance without overly constraining design.

The project team brainstormed an initial set of fifteen scenarios describing information needs that could be satisfied by data from the Federal statistical agencies and then refined them with feedback from statistical agency personnel. We identified a subset of four scenarios that were suitable tasks for the participants in this study. (Further detail on the scenarios may be found at http://ils.unc.edu/govstat/papers/scenario_paper_nov_14_2002.doc. The scenarios chosen were:

1. For one of your college classes, you've been asked to investigate the economic status of a particular county in Nebraska for a group project on the state of

https://doi.org/10.23106/dcmi.952107362

Nebraska. You have been assigned Sheridan County. Using the following websites, find 4-6 economic indicators (e.g., unemployment rate, consumer price index), and compare Sheridan County to the state of Nebraska and to the United States as a whole. Try checking all the websites to see how their data compare.

2. During your exploration of the economic health of Nebraska, you discover that the unemployment rate, total unemployment, employment and labor force statistics for the most recent time period (currently August 2002) are not the same on the BLS site and the Nebraska site. Try to find out what explains the difference in the numbers.
3. You are a social activist in the Raleigh-Durham area of North Carolina and have become increasingly concerned about urban sprawl and the loss of rural areas for farming. You are looking for statistics both for the change in amount of farming lands and farming income (market value of agricultural products) in Orange, Durham, and Wake counties. What has been the change in the amount of farmland in these counties since 1992? Has there been a comparable change in income from farming? How does the change in farmland and farm income in the Raleigh-Durham area compare to the change in farmland and farm income across the nation as a whole?
4. You are considering relocating a soybean crushing plant to either South Dakota or Nebraska. You want to compare the two locations based on the availability of soybeans within the state and their cost, and energy costs, specifically natural gas and electricity. Please provide a short summary of how they compare.

We scheduled interviews with fourteen expert users and nine student users. The expert users were a mix of users from within the statistical agencies and users from outside the agencies whose work depends on data from the statistical agencies. For the expert users, we attempted to match the tasks completed to the user's domain expertise. Each participant completed 1-2 of the tasks, and we used think-aloud protocols as well as follow-up interviews directly following task completion to gather data. For each task, we provided the participant with a web page containing the text of the task as well as links to four or five sources of data that would satisfy the task. Participants were not allowed to use a general purpose web search engine, such as Google, to find information relevant to the task, but they were allowed to follow links and use website-specific search functions once they navigated to the agency sites. The ordering of the links on the task pages were varied in order to reduce the effect of link position on the participants' decisions about which sites to use.

3. Results

In analyzing the results of our study, we were interested in what we saw as the stories the data had to tell us about the interaction between metadata and users' difficulties in finding, integrating, and using data from the Federal statistical websites. We found that there were a number of different themes that emerged; we discuss each one briefly and provide a few examples.

3.1. User Knowledge

Each person has available to him or herself a set of knowledge, expectations, and assumptions that can be brought to bear during an interaction with any information system. In this study, we identified types of personal knowledge utilized during interactions and the ways in which they helped or hindered (if the knowledge, expectations, or assumptions were incorrect).

Study participants articulated various aspects of their knowledge as they moved through their assigned tasks. One type of knowledge employed was that of the domain represented in the task. For example, one expert participant was comparing soybean prices in two states over time. As an expert in crops, he explained that he would gather about ten years of price data because beyond that fluctuations in inflation would explain more of the price differences than actual crop prices. The student participants, on the other hand, showed very little awareness of the need to gather data over time at all, preferring to try to find one number that would satisfy their information need.

Closely allied to domain knowledge was knowledge of how to approach a particular task. Expert participants who were asked to compare prices or regions generally pointed out that one should employ trend data to make sure that any given point estimate was not an outlier. In fact, one participant, who started with a "quick and dirty" point estimate, later found with a trend comparison that the one year he had used in his point comparison showed a reverse pattern to all the other years. For the task involving economic data for a county, state, and the United States, one participant indicated that there were two approaches she could take. She could either pursue the answers by geography by looking for the geographic entities to see what economic data she could find, or by starting with particular economic indicators and seeing whether she could get data at certain levels of geography.

Task and domain knowledge were often intertwined with general statistical or statistical agency knowledge. Participants provided information about their knowledge of particular surveys (such as which estimates are produced from which survey efforts, or the frequency with which statistics are released for distribution), and knowledge of aspects of survey and census processes and outcomes. For example, several expert participants completing the economic status task commented that when one wants county-level data, one generally needs data from censuses, since survey samples are too small to provide good estimates at the county level. Other aspects of statistical

https://doi.org/10.23106/dcmi.952107362

processes, such as the role of averages and rates, and concerns about how to assess the quality of the estimates, were also noted by the participants.

Expectations or assumptions that the user might have about the availability of information, the currency of that information, and other aspects of the systems and the information contained therein (such as how to navigate around the site; how to use tools on the site, such as databases; etc.) form another important component of the toolkit the user has at her disposal. The expert participants generally had a much better sense of what the most current data in a particular area would be because they could draw on their knowledge of the particular surveys with which they were familiar. The student participants, on the other hand, had no sense of the schedule of the various surveys and censuses, and so had no basis for knowing if the numbers they found were the most recent.

Articulations of expectations about system aspects frequently co-occurred with comments about specific system features. A frequent comment was one such as "I'd expect to find this information under this header information." Participants also had expectations about how navigational links and search functions would work.

The set of user characteristics described above were utilized throughout the tasks. They were part of the participants' processes of orienting to the task, developing strategies for completing the task, performing comparisons, searching, and navigating. Knowledge of how information was structured ("I can go in geographically or by economic indicator"), or which agency might provide certain information ("Unemployment numbers come from the Bureau of Labor Statistics") provide examples of using knowledge to strategize about the task.

User characteristics came to the forefront when participants were engaging in comparing behaviors. When comparing statistics with different dates, respondents used knowledge of what the latest available numbers were ("This is the latest number I can get for this so I'll use it"), the domain (see previous example on crop prices), and other knowledge. Statistical knowledge, such as how and when one can compare index numbers, was used by some participants working with the Consumer Price Index. Knowledge of survey processes and agencies was used to compare data from different sources. Several participants found it difficult to compare some numbers available from state-level economic websites with national numbers because of lack of knowledge of the actual source of those data.

This section has identified the types of knowledge, expectations, and assumptions users bring to a system interaction of this kind. It has also conveyed a sense of the use of these characteristics throughout the interaction. It is important to remember that interactions are just that, interactive, and each user responds in unique ways given his or her characteristics and processes of interaction. It is also important to keep in mind that user knowledge and expectations influence the user's experience with each of the other categories we discuss. Thus many of the examples we provide in the following sections relate to the knowledge the user brings to the table.

3.2. Surveys and Statistics

We identified several important dimensions of statistical information sources from participants' commentary and actions.

Censuses and Surveys: Specific and General Knowledge. Censuses are more likely than surveys to provide data for smaller geographic units (counties and smaller). This is due to larger numbers of cases, which enable estimates to be made for smaller units. Surveys, even large ones, will not have sufficient numbers of cases to make reliable estimates within smaller geographic units.

Additionally, censuses tend to be conducted less frequently than surveys (often only every five or ten years), which means that a "current" number from a census can be several or many years old. Generally, surveys may be conducted yearly, on a quarterly basis, or even monthly, providing statistics that are often only one month to a year behind the current date. Expert participants in our study often articulated their knowledge of specific survey and census cycles, which enabled them to judge what the most recent statistic might be. They also often knew which survey produced a given statistic and could navigate to a particular number via links organized by survey efforts. For example, one participant knew that some monthly unemployment numbers come from the Current Population Survey (CPS) and found those numbers by locating the CPS program information.

Examining specific statistics often entailed use of knowledge of the specific source. On the Census site, one can find 1999 and 2000 numbers for household sizes and money income, for example. Knowing that these numbers are generated from the 2000 Census helps explain why they are not more recent. Knowledge of the surveys also enabled some participants to determine when the next number might be available. Taking this one step further, at least one participant was knowledgeable about the publications that were generated from a given survey effort and wanted to find the relevant numbers from that survey by finding the publication that reported on that survey.

Averages, Rates, and Absolute Numbers. Some statistics are reported as absolute numbers, others as some normalized number such as an average or a rate. When making comparisons, it is often more valuable to compare normalized numbers, which will remove certain effects (such as population size or geographic area). For example, one participant commented that *manufacturers' shipments* was not a good economic indicator because it is a raw number, rather than normalized for population size or some other factor. This makes it difficult to say how one state compares to another. Some participants in the study ported numbers into spreadsheets to perform normalizing processes (such as creating a yearly average).

https://doi.org/10.23106/dcmi.952107362

Indexes. Indexes form a special class of normalized number, with the result that some comparisons are not appropriate (such as comparing the Consumer Price Index for one city with the Consumer Price Index of another).

Units. A related problem deals with understanding the units in which a particular statistic is reported. Difficulties here include undefined abbreviations for units (such as bu for bushel at the National Agricultural Statistics Service), unclear labeling (such as the label "April 1999" being used both for statistics that cover just that one month and statistics that report the year to date figures up to that month), and differences between agencies or even programs within agencies as to how an aggregate unit is defined (e.g., the Midwest region includes different states in different agencies).

Seasonal Adjustment vs. Non-Seasonal Adjustment. In our economic tasks, participants were faced with some statistics that had been seasonally adjusted and some that had not. In order to compare numbers appropriately, they needed knowledge about what seasonal adjustment entails and whether one can compare seasonally adjusted with unadjusted numbers.

Definitions and Terminology. Some participants were quite careful about noting the definitions of the variables that underlie the statistics. For example, one participant was careful to get the unemployment number at the national level for the non-farm, civilian population to match a state-level unemployment number. Knowing which classification scheme has been used to code data also is important in making comparisons. Statistics reported via the North American Industrial Classification System will not map directly to those reported with Standard Industrial Classifications.

Revisions of Numbers. Statistical agencies sometimes go back and revise published statistics, either because they have gotten additional data or because an error has been found. Several respondents commented on the need to know the "history" of a number, particularly if one needs to return to the same data set.

3.3. Interpretation of Information

Once statistics are located, users are in the position of interpreting their meaning. In our study, we identified a variety of interpreting behaviors involving understanding one statistic or integrating multiple statistics.

Interpreting One Statistic. In previous sections, we have indicated the variety of knowledge that participants brought to the table in support of their activities. Most of these types of knowledge are brought to bear on understanding a particular statistic. Expert participants noted carefully the date of the statistic and the units in which it was reported. They were alert to abbreviations (such as the use of *p* to indicate a preliminary number) and footnotes. In the economic scenario, one of the questions often asked by expert participants when looking at some of the statistics was whether they were seasonally adjusted or not. This information is often in footnotes, and expert participants tended to be aware of the existence of footnotes and often checked them. This was in stark contrast to the student participants, who often were not aware of the concept of seasonally adjusted numbers. They also were not aware that important and useful information was to be found in the footnotes, and generally ignored them.

One trouble area for participants was identifying the source of a statistic. On the Federal sites, they knew that the agency that sponsored the site with the relevant statistic was likely to be the agency that had generated the statistic. The only instance where there was a problem at the Federal level was in the use of Mapstats on Fedstats (http://www.fedstats.gov), which brings together data from several agencies. There was one instance where the Mapstats "version" of the statistic did not match that of the sponsoring agency's site, and this led to some confusion as to why that would be the case. It was the result of Mapstats' updating cycle being offset from that on the agency's site. Finding source information was more challenging on state websites. While participants could often find information about the source, it was not clear whether it was the proximate or ultimate source. For example, the Nebraska economic website indicated that its economic statistics came from the State Department of Labor, but it was not clear if the state had done its own survey or used data from the Bureau of Labor Statistics.

Integrating Multiple Statistics. Along with the types of interpretations involved with individual statistics, participants were often asked in the tasks to "put statistics together" to create larger packages. For example, in the economic task, participants were asked to find four to six economic indicators to express the economic status of a county. In these integrating situations, users performed a variety of comparisons and several other actions such as manipulating statistics or substituting one for another.

Comparisons. Once respondents had identified a set of statistics, they were faced with statistics that had different dates, defined for different geographic entities, from different sources, with different definitions of the variable in question. Respondents had different strategies for resolving these discrepancies.

Different Dates. Two strategies were identified here. Some participants attempted to find a set of indicators that were all from the same time period, which meant that they did not always use the latest number if they could not get all the statistics to be that current. A second strategy was to think through what could have occurred during the time period of the difference to see whether one still felt that the comparison was appropriate. For example, in the soybean crushing scenario, one participant commented on prices of soybeans varying in terms of whether it had been a good year for agriculture. Finding trend data was one way in which some respondents tried to understand what had happened over time.

Different geographic entities. There are two aspects to this comparison. The first is finding the same statistic across a

https://doi.org/10.23106/dcmi.952107362

range of geographic entities (such as nation, state, county, city). The second is finding a set of statistics, all for the same geographic entity. In the first instance, participants sometimes changed the statistic of interest if they could not find it at all geographic granularities of interest. At least one participant substituted the statistic from a larger granularity for the smaller one (he used the Consumer Price Index for the Midwest as a stand-in for the Consumer Price Index for Nebraska). In the second situation, the behaviors were the same. If a given statistic was not available at the given geographic granularity, participants often moved on to choosing a different statistic. The problem is complicated at the sub-state level, where respondents occasionally had to distinguish (and choose from) statistics that might be available at the county level (e.g., Yakima County), the Metropolitan Statistical Area level (Yakima MSA), or the city level (Yakima city). No one attempted to determine what the overlaps were among these closely related geographic units.

Different Sources. Many topics cross statistical agency boundaries, with the result that participants may find statistics at multiple sources. Our tasks asked them to use specific sites, and thus we do not see the full range of possible comparisons here. In fact, most participants seemed willing to take statistics from any of the sources, not commenting on whether the quality or other aspects might be different. Some respondents tended to favor particular sources with which they were familiar so that they had fewer comparisons to make. In general, the source itself was not commented on as much as the date of a particular statistic, the process by which the statistic was generated (survey or census), and the variable definitions.

Different Variable Definitions. Most of our scenarios did not lead participants to find multiple statistics for a given concept. The one scenario in which this situation arose was in the economic scenario, where respondents could find an unemployment number from the 2000 Census on http://census.gov and current numbers produced via the Current Population Survey on http://bls.gov. The expert participants who found these similar statistics commented on the source of the statistics (census vs. survey) and were aware of the differences in methodology (and thus variable definitions) that generated the conceptually similar statistics.

Manipulations. Another integrating behavior demonstrated by some participants was manipulation of the found statistics. One participant cut and pasted relevant numbers into a spreadsheet for two purposes: to generate yearly averages from monthly numbers and to display data from two states graphically. Other participants calculated averages or percent changes from given statistics in order to make comparisons.

3.4. Date Issues

One issue for the study participants concerned the timeliness, or currency, of the statistics they were attempting to use. The amount of difficulty that this issue caused for study participants was related to how much prior knowledge they brought to bear about particular agencies, their surveys and censuses, and the frequency with which those surveys and censuses are conducted. The expert participants who were familiar with the agency data could quickly determine if the particular statistic they were looking at was the most recent based on their prior knowledge.

While the expert participants were more cognizant of the issue of currency when dealing with one statistic, both the expert and the student participants encountered difficulty with trying to reconcile the currency of different statistics. This was especially an issue for the economic status task. Participants tended to view the most recent date of the first indicator they encountered as the “baseline,” and then attempted to find other indicators whose dates matched that baseline. This was difficult for the participants, because the agency websites are not designed for users to navigate through the statistics in this fashion. Information about the intervals at which particular statistics are reported, and how quickly they are disseminated to the website after the reporting interval has passed, is not readily available. Thus, there were situations where a participant would find the first economic indicator reported for the previous month at the Federal level, but then either a) be unable to find a corresponding figure for the same month at the state or county level or b) be unable to find other indicators at the Federal level for the same month. This led to backtracking, as the participant would attempt to find a set of statistics that matched in terms of time period.

3.5. Geography

Another cluster of themes that emerge from the data we gathered from study participants concerns the very important and surprisingly complex role of geography in participants’ usage of statistics.

One issue with geography concerns the user’s prior knowledge of geography and the extent to which the user needs to possess a knowledge of geography to navigate to the data she wants. In many instances, the study participants were confronted with a map of the United States that had no labels on the states. There was an assumption that the user would know which state to choose, but many of the participants who completed the soybean crushing plant task could not identify which states were Nebraska or South Dakota on the map of the United States. Similarly, there were state maps that did not have county labels, or points representing major cities. This made it very difficult for the participants who were trying to navigate to information about Yakima, Washington, because not knowing much about the state they had no idea which county they should choose.

Another set of problems can be grouped under the heading of geographic granularity. We define this as the unit of geography to which a particular statistic applies; so

https://doi.org/10.23106/dcmi.952107362

for instance you can get a figure for the Consumer Price Index for the nation as a whole, for some region of the nation (such as the Midwest), or some particular state (such as New Jersey). But you cannot get a figure, in most cases, for the Consumer Price Index at the level of a county within a state, a city, or a Metropolitan Statistical Area (except for some large urban areas). Geographic granularity caused problems for participants both in instances where a particular statistic was simply not available at some geographic level (like the Consumer Price Index for a county, mentioned above), and in instances where the participant could not readily discover whether or not she could expect any particular statistic to exist at a particular geographic level.

Geographic granularity issues intersect with currency issues, especially when trying to vertically integrate statistics from the Federal level down to the level of counties and cities. The date of the most recent statistics available at the Federal level may not be the same as the date of the most recent statistics available at more specific geographic levels. This was frustrating to both expert and student participants, who were disappointed that there was not way to determine at a glance if a particular statistic existed for a particular combination of time period and geographic level. When participants could not find statistics at the desired geographic level, they were unsure when it would be appropriate to substitute a statistic form a different geographic level. For example, when one participant could not find the desired statistic for a particular state, he used the statistic that applied to the region in which the state was located, deeming that to be close enough.

The geographic substitution problem was compounded in situations where one name is given to more than one geographic entity, as in the case for the city of Yakima Washington, Yakima County in Washington, and the Metropolitan Statistical Area named Yakima. In cases like this one, many participants would use these three kinds of entities interchangeably, and seemed not to notice that they were looking at data from the MSA when the task specified that they should be looking for data for the county. In the cases where participants did notice the discrepancy, they were not always sure what to do about it. These participants reasoned that in these cases the data for the city, county, and MSA might reasonably be considered to be interchangeable, but since they were unclear on exactly what the differences were between these groupings, they could not be sure.

3.6. Navigation

Many of the problems participants faced in finding the data they needed for the tasks they were given involved the navigational paths and aids provided by the agency websites. One of the biggest issues deals with labeling of navigation links. These labels are often representative of groupings that make sense from the agency standpoint, but do not necessarily match the way the user would group the concepts. One example of this appeared repeatedly in the soybean crushing plant task. One of the best sources of information to complete this task was in the Agricultural Statistics Data Base at the National Agricultural Statistics Service (NASS). At the top level of this database, there are links that correspond to the major categories of crops. Many participants used this database in the hopes of finding information about soybean yields in the states of South Dakota and Nebraska. The top level categories in this database are Grains; Hay; Oilseed & Cotton; Potatoes, Dry Beans, & Hops, Sugar & Mint, Tobacco; and Vegetables – Fresh and Processed. Faced with these choices, most participants (including, interestingly enough, the agency expert participants from NASS) picked the “Potatoes, Dry Beans, & Hops” category. On finding that soybeans were not a part of this category, most participants were stumped. Some continued to try the other choices until hitting upon the correct category, which was “Oilseed & Cotton”, while others assumed that since soybeans were not listed in the “Potatoes, Dry Beans, & Hops” category, there must not be any information about soybeans in the database. These participants thus had a much harder time finding the information about soybean production and yields, as the other resources provided were not as good..

This labeling problem was also evident in participants’ use of A-Z lists, which are a feature of many of the statistical agency websites. Again, the labeling of concepts in these lists is often reflective of the data producer’s view of categories rather than the view of the data consumer, so these lists could be very frustrating for the participants. Participants were also unclear about how comprehensive the A-Z lists were supposed to be in terms of the range of concepts about which information existed. If a participant could not find an appropriate category on an A-Z list, she was unsure if that meant a) the category was labeled with an unexpected name; b) the category of information she was seeking existed on the website but was not covered in the A-Z list; or c) that category of information simply did not exist on the website. How the participants reacted to this problem varied depending on how strongly they assumed that the particular concept they were looking for did, in fact, exist on the site, and how strongly they assumed that the A-Z list was comprehensive. Those participants who felt certain that the concept they were looking for was there under some other name spent more time trying alternative options from the A-Z list. Those who were uncertain that the concept they were looking for was covered tended to abandon the A-Z list and go on to other strategies.

The use of maps as navigational aids also proved to be a mixed experience for participants. Many participants expressed satisfaction that a map was provided, but how useful the map was to them was directly related to the way in which it was presented. Participants wanted maps that were large, well-labeled, and clickable. The participants could easily find a state in which they were interested on a

https://doi.org/10.23106/dcmi.952107362

US map with the states labeled, but still had difficulty finding a particular county on a state map with the counties labeled, because they had no prior knowledge in many cases of what part of the state would contain the county in question. Participants in these cases expressed a desire for a text alternative presented simultaneously with the map that would allow them to find the particular unit they were looking for from a drop-down list. Maps that were not clickable and not labeled were seen by participants as not being useful.

Some participants, especially the students, seemed more comfortable with the idea of searching for information rather than trying to navigate to it using the options given on the websites. In fact, a number of student participants asked (to the point of nearly pleading) if they could use Googleto find the information specified in the tasks. For the purposes of this study we did not allow participants to use a general web search engine, but we did allow them to use whatever search functions the agencies provided on their sites. Participants had mixed results with searching for a number of reasons. First, on many sites the search engine was indexing only a subset of the pages on the site. Second, when a participant's search would return no results or irrelevant results, the participant was again faced with being unsure whether a) the results reflected differences in agency labeling of concepts vs. user labeling of concepts; b) the results reflected gaps in coverage of the site content by the search engine; or c) the results reflected the actual lack of the sought after information on the site.

3.7. Information Layout

Related to the issues of navigation are the issues of how that information is laid out and presented to the user. In a number of cases, the participants would navigate to a document that contained information relevant to the task, but would not notice that the information was there because of the way it was presented. For examples, many tables at the Bureau of Labor Statistics and other agencies are presented in PDF format. These tables are often many pages long, and often groups of related tables are presented within the same document. This is not always obvious to the user; she will be presented with the title of the first table in the document, see that it does not contain the information she is seeking, and give up and go elsewhere. All the while, the information she seeks does exist in the document, it just lies within a table that starts on, say, p. 50 of the document.

A related problem occurs when the column headings for a table are shown only at the top of the table, and the rows spread across multiple pages. This requires the user to scroll down to find the appropriate row, realize that she no longer knows which column she was using, scroll back to the top to determine the appropriate column, and then scroll down to the appropriate row again. This is frustrating for users, and is also prone to error, as it is quite easy to lose track of the column while doing all of this scrolling.

Another related problem is the way linking is handled within the documents. For example, one of the participants who was trying to determine whether a set of statistics was seasonally adjusted or not used a link to get to the footnotes for a table. She clicked on footnote #2, which did not contain the information she needed. Footnote #1, however, did contain that information – but the way footnote #2 was displayed to the participant did not allow her also to see footnote #1. So she missed the piece of information that would have answered her question.

These kinds of issues are representative of the pitfalls of building a dissemination model for the Web that is closely tied to the print dissemination model. PDF files are a convenient way to present information formatted for print in an electronic format, but they fail to capitalize on the presentation advantages, or acknowledge the presentation constraints, that are inherent in a hypertext environment.

3.8. Terminology

Use and Understanding of Technical Terms. Experts need specialized terminology, but terminology can also serve as a barrier to finding or understanding the information. Our expert participants often recognized that some terms were problematic, but could also whip off a series of acronyms without seeming to realize what they were doing.

People sometimes try to guess or infer meaning from the term. One participant said,"total non-farm I'm sure it includes total manufacturing. Total non-manufacturing would be everything but the manufacturing." In this case, it works, but in other cases (such as the term "seasonal adjustment"), it is not necessarily as intuitive.

Terminology problems seem to arise when people are looking for information, when they must choose which statistic or table to use, or when they are trying to compare statistics. For example, in order to find information about soybeans, the user must look under oilseeds. Seasonally adjusted vs. unadjusted data is a problem both in choosing which to use, and also in deciding whether one can do a "mixed comparison", comparing an adjusted number with an unadjusted one.

Statistical terms such as seasonal adjustment are one source of problems,. Domain terms also cause trouble, and it is sometimes difficult to tease out where knowledge about terms can be differentiated from knowledge about the processes and entities in the domain. In the soybean task, for instance, the amount of soybeans on hand is measured both on the farms, and in the elevators (off-farm). These represent two different stages in the life cycle of the soybean, and therefore play a role in availability. Similarly, the fact that most natural gas is off-system, and what that means, is generally hidden from the user except as a footnote.

How much does someone need to know? Enough to find and use the information, but probably not as much as

https://doi.org/10.23106/dcmi.952107362

an expert. For example, end users probably do not need to know or understand the formula for seasonal adjustment.

Some terminology problems are clearly artifacts of the history or purpose of the survey or census. Even when the agency recognizes the problem, it can be a difficult one to address. The Energy Information Administration's use of "average revenue per kilowatt hour", which essentially means "retail price", is an example.

Mappings Between Language for General Purpose (LGP) and Technical Terms, Ordinary Concepts and Related Specialized Concepts, and Synonyms. The crux of this facet is that much of the information describe by the agency statistics is based on concepts and processes that are familiar to many users, at least to some extent. Jobs, utility costs, and farm production, for example, are not really hidden from sight. Users may use different words or phrases to describe them, or have slightly different definitions of the concepts, but a level of "regular life" understanding can help users get started finding and using the information *if* they can make the leap from their words to the agency words. A common example is using *Consumer Price Index* to refer to *inflation*, but the EIA *average revenue per kilowatt hour – retail price* is another one. On the other hand, there are times when the general level of understanding can be a pitfall, leading someone to believe she understands the concept when she really does not. The difference between a city and the MSA by the same name is an example.

Labels as Signposts, or Use of Terms in Prominent Places. This facet is closely related to the theme of expectations and finding or recognizing information. Terms used in links, table headers, tables of contents, etc., play important roles in users' finding and recognizing the information they need. When users do not understand the terms, there is a risk that they will overlook or ignore useful information, on the one hand, or that they think they have found something useful when they have not on the other. When the label is isolated from the table itself (e.g., in the table of contents or list of links), there is little additional context that could provide hints that the user is (or is not) on the right track. For example, one participant found the link "Electricity Prices", which sounded useful. When he arrived at the table, it turned out to be "gas sold to electric utilities" instead. An expert participant pointed out that the EIA site stores historical data under "short-term forecasts", which is not intuitive to most users. Another example shows an attempt to solve the mapping problem between LGP and agency language. The link is "Petroleum Consumption", but the table is labeled "Petroleum Product Supplied". The expert participant pointed out that what is missing is a note that says these two phrases mean the same thing – otherwise the user might think the link led to the wrong place.

Use of Terms as Sign of Expertise or Familiarity. One sign of expertise is in the use of technical terms, acronyms, etc. Not only does this generally signal familiarity with the domain concepts (although this does not guarantee a deep or correct understanding), it also indicates better recognition of desired (and undesired) information, even on a site with which the user is unfamiliar. For example, someone who understands seasonal adjustment is more likely to understand why some data is offered in adjusted and unadjusted form, and possibly to make the appropriate choice between them, even if they have never used a parrticular agency's site before.

Ambiguous Terms. We use "ambiguity" here to mean "defined differently", which is different from the common *bank – river/financial institution* example. Here, a term is usually naming approximately the same concept, but the details of the definition are different. Ambiguity occurs between LGP and agency definitions, between agencies, and within agencies. "Sector" is a good example of this; it is variously used to express public vs. private, manufacturing vs. service, and commercial vs. residential, among others. Someone who has expertise on a particular survey knows the definitions of the variables, and may also know differences between definitions for various surveys. But the differences are not generally highlighted in a systematic way. At best, a user would have to read the definitions of a variable used by each agency or survey, and infer the differences herself.

Definitions Changing over Time. Related to the issue of ambiguous terms, operational definitions of a term may change over time. The MSAs are one example, as are the definitions of race and ethnicity. One participant thought there had been changes in definitions of underemployment, but could not remember the details (and could not easily find documentation on it).

Definitions Available to the User. For many reasons, definitions of terms are not easily available to the user. One expert participant had trouble finding definitions for MSA and PMSA, even though he knew they existed on the site. Many times, definitions are given in footnotes, but people often do not look at footnotes, or have difficulty finding them. In other instances, the footnotes or appendix of a paper publication were never linked to each table or occurrence of the term, or perhaps were never put on the Web in the first place. A couple of participants commented that this is the sort of thing the Web ought to make easy, but it has not been carried through.

A related problem is the nature of the definitions. Many agencies do a good job of aiming definitions at non-expert users, but others give only highly technical definitions, which incorporate additional (undefined) terms, equations, and complex language.

4. Integration Activities and Challenges

In addition to the facets discussed in Section 3 above, we also discovered a number of insights about the nature of the activities that make up the process of integrating information on the part of the user, as well as a number of barriers to that integration taking place successfully.

https://doi.org/10.23106/dcmi.952107362

The most important integrating activities we observed include the following:

- Making comparisons
- Noting discrepancies (between data, in presentation approach, etc.) and/or asking what the difference is due to
- Manipulating statistics (e.g., mathematical, exporting to spreadsheets)

Of these, the comparison activity appeared to be the core integrating activity. We identified a number of types of comparisons common across our participant pool; these types include:

- Comparison across geographic units
- Comparison when there are definitional differences across concepts and variables
- Comparison across units of time
- Comparison across different sources (websites, surveys, censuses, reports, etc.)
- Comparison across index values

We also identified a number of barriers to the successful integration of statistical information. These barriers include:

- Lack of definitions or source information
- Lack of user knowledge of appropriate strategies (e.g., using time series data, types of calculations to perform)
- Lack of user knowledge about usage of index values, statistical activity purpose and approach
- Interface design problems (such as scrolling row and column headers)
- Inconsistent data across sources
- Inconsistent interfaces
- Inability to determine whether data wanted for comparison are available
- Lack of domain knowledge
- Lack of knowledge of how to handle domain terms such as inflation, seasonal adjustment
- Terminology differences

5. Metadata

The ultimate goal of our analysis was to map the different themes we discovered to specific kinds of metadata elements, develop an XML schema to contain them, and explore the implications for the kinds of systems and interfaces that should be built to house and use that metadata in a way that will be effective for the end user.

The first step was to identify common user problems for which metadata might be useful. The problems we identified are as follows:

- Mapping of agency terms to user terms
- Definitions of statistical/survey terms
- Comparability of statistics
- Help with finding and interpreting statistics
- Information about recency of statistics, update schedule, when updates available on website
- For specific statistics, geographic levels at which that statistic available
- Navigation by means of something other than large lists of text links
- Column headings that are always visible

5.1. Metadata Elements

From the problems above, we have started to derive a short list of elements (to be expressed in XML) that can enable discovery and use of statistical information. These might include:

EntDscr (Entity Description)
---EntGroup (Entity Group)
---EntType (Entity Type)
---Titl (Title)
---IDNo (ID Number)
---AuthEnt (Authoring Entity)
---ProdStmt (Production Statement -- Marked-up Document)
---Producer (Producer -- Marked-up Document)
---ProdDate (Date of Production -- Marked-up Document)

StdyDscr (Study Description)
---StdyInfo (Study Scope)
---Subject (Subject Information)
---Keyword
---Abstract
---SumDscr (Summary Data Description)
---TimePrd (Time Period Covered)
---CollDate (Date of Collection)
---GeogCover (Geographic Coverage)
---GeogUnit (Geographic Unit)
---AnlyUnit (Unit of Analysis)
---Universe
---DataKind (Kind of Data)

VarDesc (Variable Description)
---Var (Variable)
---Labl (Variable Label)
---Concept
---Qstn (Question)
---Key (Range Key)
---Range (Variable Range)

TermDesc (Terminology/Ontology Description)
---Concept
---Term
---Present (Presentation)

This initial set is drawn partially from a set developed by the Data Documentation Initiative (DDI) [3] as well as from an ontology element set, the DAML+ OIL [4] and a classification element set, the Neuchatel Terminology Model [5]. We found a critical need for metadata describing aspects of time/date and geography. The existing elements in DDI need expansion to express these aspects. We propose including an element for date as well as an attribute for periodicity on the Entity Group (<EntGroup>) element. This pairing enables users to find statistics for both a given time period as well as determine the frequency with which a statistic is updated. We take a similar approach to geography, providing both a Geographic Entity element as

https://doi.org/10.23106/dcmi.952107362

well as a geographic entity attribute, again enabling both requirements for geographic information to be satisfied.

5.2. Metadata Challenges

This element set does not yet address all the metadata requirements for the SKN but does enable us to prototype several discovery and interpretation tools. These prototypes are described in detail in the papers available from our project website, see http://ils.unc.edu/govstat/papers.html. For example, user difficulties with terminology and understanding statistical and domain concepts can be addressed with an enhanced glossary tool. The Statistical Interactive Glossary (SIG) is a prototype designed to provide users with definitions and examples in context and in a variety of forms [6]. SIG and its back-end ontology require metadata elements for concepts and contexts, as well as ontological relationships. In future years of our project, we will continue to expand the element set to address further challenges.

One area is creating a mapping of agency terms to user terms. We will first need to refine our knowledge of what terms users are employing to refer to agency concepts, and we will need a better understanding of how the agencies themselves are using their terms [7]. We will also need to map terms across agencies, as they may be used differently by the different agencies in different contexts.

A particularly challenging area of user support that can be enabled by metadata is that of helping users strategize about how to find and use statistics and helping them make appropriate comparisons. At the first level, we must have the metadata that strictly defines variables and can maintain linkages between variables that are defined on the same units. Metadata on variables is well-developed in the statistical agencies. (See the DDI [3] for detail on what metadata is retained about variables.) Beyond that, providing users with in-context strategic help such as when to look for survey vs. census data, or facilitating comparisons requires both new types of information not currently produced by statistical information and a technique for managing them.

6. Future Work

We are continuing to identify the sources of metadata available to us (particularly automatically) that are relevant to the problems described in this paper. In some cases this is fairly straightforward, but in other cases not so clear. There is a great deal of technical information about how particular surveys and censuses are conducted, and we can draw on such technical documentation to help users understand the survey/census process. The agencies have definitions of the variables and units used that would be helpful for users to understand what a column in a table really means. In the case of dates and geographic entities, it is less an issue of availability than it is of organization, navigation, and layout, as discussed above. In addition, metadata concerning release schedules for surveys/censuses can be made more visible to users.

Having identified some common problems and the kinds of metadata that could help alleviate these problems, the next step is to incorporate specific metadata into a variety of tools that are under development by other members of our project team. For this effort we are drawing on the findings from previous grant work in this area [8]. We also need to determine when these tools are appropriate to particular user situations so that we can provide "just-in-time" help modules that do not remove the user from the context of her work. The most difficult aspect of this will be trying to anticipate what the user needs based on her expectations and prior knowledge, which will necessarily be different for each user.

Acknowledgement

This material is based upon work supported by the National Science Foundation (NSF) under Grant EIA 0131824. Any opinions, findings and conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

References

[1] Carroll, J.M. (2002). *Making Use: Scenario-based Design of Human-Computer Interactions.* Cambridge, MA: MIT Press.
[2] Rossen, M.B. and Carroll, J.M. (2001). *Usability Engineering: Scenario-based Development of Human Computer Interaction.* Morgan-Kaufman.
[3] Data Documentation Initiative. (DDI).http://www.icpsr.umich.edu/DDI/index.html
[4] DAML + OIL (DARPA Agent Markup Language + Ontology Inference Layer) http://www.daml.org/2001/03/daml+oil-index.html
[5] Neuchâtel Terminology: Classification database object types and their attributes, Version 2.0 (2003). Currently not publicly available but available from Carol A. Hert.
[6] Haas, S.W., Pattuelli, M.C. & Brown, R.T. (2003). Understanding statistical concepts and terms in context: The GovStat Ontology and the Statistical Interctive Glossary. *Proceedings of the American Society for Information Science and Technology Annual Meeting*. (to appear).
[7] Haas, S. W. (2003). Improving the search environment: Informed decision making in the search for statistical information. *Journal of the American Society for Information Scence and Technology*, 54(8) 782-797.
[8] Marchionini, G., Hert, C., Shneiderman, B. and Liddy, E.. (2001). E-tables: Non-specialist use and understanding of statistical data. Proceedings of dg.o2001: National Conference for Digital Government. (Los Angeles, May 21-23, 2001). 114-119.

https://doi.org/10.23106/dcmi.952107362