# STUDIES OF METADATA CREATION AND USAGE[1]

**Carol A. Hert**
**School of Information Studies, Syracuse University**

**keywords: metadata; survey methodology; user studies**

**Abstract**

The important role of metadata in survey development and integration processes, in quality assessment, dissemination, and other activities is becoming increasingly apparent in the statistical agencies. Agencies have also recognized that metadata is expensive to create and maintain, and that even in an ideal world not all metadata can be saved, nor is all metadata equally useful to all the stakeholders of statistical data. One approach that can provide information relevant to decision making associated with metadata repositories is to investigate how various stakeholders create and use metadata. Such user studies can highlight components of metadata that are frequently used (in particular tasks), identify metadata that are important but not yet captured in systems, and provide information about the context of usage relevant for system design. The paper provides brief overviews of three projects investigating metadata usage with specific user groups. While the studies were not directly comparable, general themes in metadata use were identified. Metadata communicates information deemed important for quality assessment and additional information/metadata may be necessary to support that task. Metadata capable of expressing contextual elements is helpful. Accessing and understanding metadata may be difficult due to metadata structure and language as well as due to its inaccessibility.

## Introduction

The role of metadata in survey development and integration processes, in quality assessment, dissemination, and other activities is becoming increasingly apparent in the statistical agencies. Agencies have also recognized that metadata is expensive to create and maintain, and that even in an ideal world not all metadata can be saved, nor is all metadata equally useful to all the stakeholders of statistical data. One approach that can provide information for decision making is to investigate how various stakeholders create and use metadata. Such user studies can highlight components of metadata that are frequently used (in particular tasks), identify metadata that are important but not yet captured, and provide information about the context of usage relevant for system design.

Over the last several years, the author, with colleagues, has been engaged in three separate projects that each explored how a particular stakeholder group created and/or used metadata. These groups were 1) expert users retrieving microdata for analyses, 2) more novice users working specifically with statistical tables, and 3) survey methodologists engaged in two specific survey efforts. Each of these investigated dimensions of metadata usage and together they provide a baseline for other studies both in terms of possible methodological approaches and findings. The intent of this paper is not to fully explicate the studies, rather it will highlight over-arching themes across them.

[1] The work reported in this paper was funded by the National Science Foundation Digital Government Initiative, the United States Bureau of Labor Statistics, and an American Statistical Association/NSF/BLS Research Fellowship.

## Defining Metadata

The most basic definition for metadata is "data about data." A more useful set of definitions in the statistical arena is provided by Dippo and Sundgren (2000) who offer:

- Metadata is the information used to interpret/use/understand the information
- Metadata describes or documents statistical data (over the lifetime of that data from survey conceptualization to data dissemination)
- Metadata includes resources and tools used in producing data (such as instruments, interviewer documentation, coding schemes)

In the domain of the statistical agencies, one can identify a narrow definition of metadata that focuses on information about the data (e.g., variances, response rates, response categories) and information used to produce the data (such as data collection instruments, instructions, technical documentation). There is also a broader sense of metadata that considers as metadata more information about the data and processes used (for example by including information on results of cognitive tests of instruments, perhaps even the videotapes of respondents) and by extending the survey cycle through the dissemination process (thus incorporating metadata related to website dissemination and retrieval etc.).

These definitions continue to be discussed and agreement on a standard definition seems unlikely. What the community does agree upon is that statistical metadata consists of both structured and unstructured information; is highly varied in its usage through the survey lifecycle, tends to support some tasks better than others, is expensive to create and maintain, and that agencies may not be capturing all the necessary metadata to support the diversity of tasks either within an agency or by the users of an agency's data[2].

## A User Perspective

The situation above presents an argument for a user-centered perspective. When experts can not agree on what metadata should be captured, in what formats, or at what expense, examinations of users/stakeholders can provide insight into which metadata tend to be used; what formats are preferred, and important metadata characteristics.

The user-centered perspective is not monolithic, instead it might be viewed as a spectrum of theoretical, methodological, and analytic perspectives that privilege user perceptions and constructions of reality. Thus a researcher seeks to understand how the users define their realities and how they act within these realities. This paper cannot do justice to these; however, it is important to note the diversity, and establish that user-centered approaches use diverse methods and generate findings that are generalizable to greater or lesser extents. The primary rationale for the employment of user-centered methods

[2] There are a variety of metadata efforts including the Inter-University Consortium for Political and Social Research's (ICPSR) Data Documentation Initiative (DDI); an International Organization for Standardization metadata repository standard (ISO/IEC 11179); an UN/ECE Work Session on Statistical Metadata (see for example: http://www.unece.org/stats/documents/2000.11.metis.htm) and various agency-level efforts.

typically is not to seek generalizability but to acquire a richness of understanding. Such richness provides the basis for variable conceptualization and operationalization as well as for proto-theories that might be tested more formally.

The three studies reported on here all began with the assumption that by investigating people's interactions with metadata, insights would be gained as to how to provide more useful metadata in more useful formats. Additionally, in two of the studies, the definition of metadata was extremely broad so that new types of metadata might be identified. The populations, methods, and analytic approaches varied across the studies (as per the discussion above).

**Study 1: Using Metadata In Microdata Retrieval Tasks**[3]

The first study, conducted in 1999, investigated how expert analysts within the Bureau of Labor Statistics (BLS) and also from several research organizations in the Washington D.C. area used metadata available within the FERRET data access tool (http://ferret.bls.census.gov/cgi-bin/ferret) to determine variables to be used in particular analysis tasks. The goal of the project was to identify which metadata were used by these analysts, barriers to use of that metadata, and the enhancements that analysts wanted. A total of 10 people participated in one of two sets of tests.

The researchers (the author and John Bosley of BLS) developed and tested a set of usage scenarios in which respondents were asked to determine variables in the March supplement in the CPS that they would use in a given analysis. For example, one scenario asked respondents to identify the variables they would use to calculate the unemployment to population ratio for several states and metropolitan areas. Choices were made from a set of variables with metadata identified by the researchers, with the help of agency experts. As the respondents worked through the scenarios, we asked them to "think-aloud" and we audio-taped their commentaries and conducted a debriefing interview after the scenarios were performed.

The study provided information on which metadata elements were used by expert analysts to determine variable choice, as well as what aspects of elements were problematic for the experts. We found that:

- Context matters. How questions relate to one another is used to understand who might have been asked the question. The experts relied on their understanding of the survey and its purpose to determine whether the variable might be appropriate.
- Variable naming conventions are used. Analysts indicated that they rely on their knowledge of variable naming conventions as a quick guide to variables
- Universe statements matter. Knowledge of the number of people or proportion of the sample a given question reaches is extremely important information
- Valid item values need to be clearly written. The way the metadata file presents valid item values was perceived as not always clear or salient.
- Standard variables and recodes are often preferred.

[3] Full details of the study can be found in Hert (1999) available at http://istweb.syr.edu/users/faculty/hert/BLSphase3.PDF

- Non-public information is used: The BLS analysts have access to some information about variables, not available to the public. This documentation is organized differently and includes an index that several BLS analysts used during the scenario task to make their decisions.
- Coding categories help analysts interpret the question. When the question might be unclear to the analysts, the available coding categories were used to provide additional information on the nature of the question.
- A variety of strategies are employed: In addition to using the metadata, the analysts reported on other activities they might use to understand the variables. These were looking at the questions in context, checking published numbers against numbers from their calculations, choosing multiple variable options, doing exploratory statistical analyses, and reading the survey documentation.

**Study 2: Exploring Metadata For Understanding Statistical Tables**[4]

The second study was part of an NSF-funded project[5] concerned with improving access to statistical tables by incorporating functionality into web-based tables and by supporting user understanding. To support user understanding, we determined types of questions users had about specific tables, then searched for answers in metadata sources. The intent was to determine 1) if metadata existed to resolve user uncertainties and 2) if extant, whether it could easily be incorporated into a variety of online help systems.

As with the previous study, our assumption was that lacking information about usage of tables, we needed to employ a user-centered approach. Rather than starting with what metadata was available and possible to incorporate into tables, we started with user questions. This study also used a broad definition of metadata as discussed earlier. In this case, any documentation (including technical documentation, online FAQ's, footnotes to tables, etc.) was included as a metadata source to be explored.

Eleven volunteers participated in the study. Each viewed a total of three tables[6] in a mix of electronic and paper formats. After an initial period in which each participant was instructed to examine the tables and think-aloud, the researchers asked a series of questions to probe for additional areas of uncertainty. Analysts reviewed the transcripts produced and extracted uncertainties, suggestions, and complaints.

---

[4] The detailed report of this study can be found in Hert, Hernandez and Armstrong (in preparation) and is available at http://istweb.syr.edu/~tables/uncertaintiespaper.html

[5] The project was entitled Citizen Access to Government Statistical Data. A variety of papers, presentations, and other information on the project is available at http://istweb.syr.edu/~tables.

[6] Participants viewed three of four tables used in the study. The four tables used were: the Economy at a Glance from BLS (http://www.bls.gov/eag/eag.us.htm), Resident Population of the United States by Age and Sex from the 1998 Statistical Abstract of the United States; Retail Gasoline Prices from the Energy Information Administration (http://www.eia.doe.gov/oil_gas/petroleum/data_publications/wrgp/mogas_home_page.html), and Estimated Average Length of Life in years by race and sex—Death Registration States 2900-28, United States 1929-1996 from National Vital Statistics Report 47(13): Dec. 24, 1998 also at http://www.cdc.gov/nchs/faststats/pdf/47_13t5.pdf

The team coded the resultant lists using schemes that were inductively derived. Researchers used data from a previous study (Hert, unpublished) to start the process of inductive or open coding (Krippendorf, 1980; Strauss & Corbin, 1990) used to develop the final coding scheme for uncertainties. Our initial set of categories was put together by identifying instances where a user had a: 1) direct question 2) expressed concern or 3) expressed confusion. A total of 170 instances were identified with a total of 106 unique instances. The scheme's reliability was assessed using a simple measure of agreement between coders. The scheme along with frequencies is indicated in Table 1.

**TABLE 1**

| Categories<br>~~Sub~~categories | Definitions/Examples | Freq. |
|---|---|---|
| **Definition needs :** | **Users ask about the meaning of something in the table.** | |
| **Meaning of Terms.** | *What does seasonally adjusted means?* | 47 |
| **Meaning of Data.** | *I'm not sure what the data cell refers to.* | 5 |
| **Meaning of Categories** | **User is uncertain of what belongs to a particular category. Ex.** *What does Non-Farm wage include?* | 18 |
| **Meaning of Abbreviations** | *What does T&P means?* | 17 |
| **Population Universe** | **User is uncertain of what population/universe data can be generalized to. Ex.** *Is it the population of the US?* | 9 |
| **Unit of measurement** | *Is this in number of persons or number of jobs?* | 1 |
| **Rationale of information** | **User is uncertain of the reasons why or how something was done, reported, computed, etc in a particular way. Ex** *Why are numbers reported differently for NY & LA?* | 28 |
| **Table structure :** | **The way the table is organized and formatted make user uncertain about the meaning of data.** | |
| **Formating, layout, and components** | *I don't understand why the numbers are in purple.* | 10 |
| **Meaning of Labels** | *I don't understand this label.* | 11 |
| **Organization of the links in webpage** | *I wouldn't expect to find Press releases first when you click on the link* | 3 |
| **Lack of information on :** | **User is uncertain of how data was collected, computed, etc.** | |
| Data collection procedures | *How was data collected? What method were used?* | 2 |
| **Sources of data** | *From where was data collected?* | 2 |
| **Computation methods** | *How were rates computed?* | 4 |
| **Comparability/Relationship of Info.** | *What is the difference between CPI-U & overall CPI?* | 6 |
| **Tool Functionality** | *Can I make a graph right now?* | 2 |
| **Updates to information** | *When was the information updated?* | 1 |
| **User uncertainty is not clear** | **The user didn't clearly explained his/her uncertainty.** | 4 |
| | **TOTAL** | **170** |

The next task was to determine whether existing metadata contained information relevant to the user question/uncertainty. The team searched for answers to the specific individual questions. Answers were sought within the table and accompanying text, related documentation (in both electronic and paper format), and in some cases, where no answer could be found, by consulting experts within the agencies that produced the tables.

Answers were found in metadata sources for most of the 106 unique user uncertainties. Sources included footnotes of the table, technical documentation for the table or series of tables (such as the appendices of the Statistical Abstract of the United States), Frequently Asked Question lists (online only) and links embedded within online tables. Very rarely

was there a direct linkage. For a question associated with the Resident Population table of the Statistical Abstract, the researcher needed to search several pages of explanatory text for the section of tables as well as appendices to the volume. For the BLS At-A-Glance tables that have embedded links, it was necessary to follow a link either to an overview of a program page or to the introductory FAQ page from which one had to choose the correct question. Only two user questions had answers directly linked.

The predominant type of question for which an answer was not found and expert help was needed was for rationale questions. Answers to these questions often require background knowledge of the history of a survey, the domain in question, etc. To demonstrate the nature of these questions, one user question about the Retail Gasoline Table was "Since it is illegal in New Jersey to have self-service gas stations, how can these be the 'Self-service prices per gallon for the country'?" The expert's answer was

> some states, New Jersey for one, do not allow self-serve. In those cases, the prices represent the only service of gasoline provided in that state. Our analysis has always shown that this is not a big price effect in those states as compared to states allowing self serve that have higher prices for full-serve vs. self serve. …The industry doesn't make an issue of it nor do we. Some states have other laws such as refiners can't operate gas stations (Maryland for one) and we don't note them either as non-refiner state stations or anything.

In addition to finding answers in the metadata to resolve user uncertainties, the team identified several issues. The first is the question of whether a user is provided with a somewhat generic answer to a question or one that specifically resolves the uncertainty. For example, one user had the question: Why are things seasonally adjusted (from the BLS At-A-Glance tables)? There is a very specific answer, but more generically, this might be considered a question that concerns a definition and a user could be provided with the definition of *seasonal adjustment*. It isn't clear which type of answer would be most efficient for both user and answer provider. A second issue is that not all answers (generic or otherwise) are easily found in any metadata source. The team found that answers may not be in electronic format at all, or may be buried within a large document thus making it difficult to retrieve.

A third issue identified is that some answers are consistent across tables, while others might only be relevant to one specific instance of a table. A question such as "Why is the 1998 statistic for urban unemployment so high in relationship to the other 1998 numbers?" would relate only to one specific cell on one specific table, while a question such as "what is the definition of seasonally-adjusted" is likely to be at least consistent at the agency level. Currently, metadata systems do not provide information on the boundaries within which a particular item is valid. It also seems that questions that demonstrate a richer domain knowledge may be harder to answer with easy to retrieve information. A question such as " Why is there no OPRG in PADD 1C and no Oxygenated in the PADD1's?" related to a gasoline table can not be resolved with simple definitions and to answer requires an additional source (A map in another document) and knowledge of how the gasoline formulations and their reporting is changing.

By utilizing a very broad definition of metadata, this study perhaps "pushed the envelope" in terms of what might be considered a metadata usage study. These users probably had little knowledge of metadata, instead answers to questions was their concern. A question that emerges then is the relationship between metadata and other types of information and when and how these sources interact to support particular tasks.

**Study 3: Metadata Creation And Use By Survey Methodologists**[7]

The third study of metadata use is a large ethnographically conducted study of how survey methodologists involved in two BLS surveys create and use metadata. The intent was to provide a description of that usage that could inform other metadata initiatives, identify additional metadata, and develop a set of recommendations and metrics related to metadata practices within this context. The two surveys, the Current Population Survey (CPS), and a new effort, the American Time Use Survey (ATUS) were chosen as they had the potential to represent two ends of a metadata spectrum. The CPS is an established survey with substantive documentation and metadata and the ATUS is a new effort scheduled for initial data collection in 2003 thus was in the process of developing metadata. The researcher was resident at the Bureau for approximately six months during which time, work practices and meetings were observed, interviews with participants conducted, and documents gathered and analyzed. Data were analyzed using inductive techniques (as articulated in Glaser and Strauss, 1967; and Guba and Lincoln, 1989).

*The Work Worlds Of Survey Methodologists*

The first dimension of the findings relates to an expression of the survey methodologists' work worlds as related to metadata. Methodologists in the two projects performed a wide array of tasks from sampling design to cognitive testing of questions to establishment of analytic procedures to determining the impacts of changes in survey design. In essence, their work revolves around establishing baselines for quality and cost-effectiveness of survey processes, assessing the extent to which those baselines have been met, and providing alternatives and strategies for improving quality. They provide these services to the organization both by doing basic research and through consultation and involvement in ongoing survey efforts. While the methodologists had a multitude of ways to express their work, uniformly, at some point, in the conversations they relied on expressing the overall goals of their work as quality assurance and improvement.

To perform their work, survey methodologists conduct and report research and participate in decision-making processes associated with the surveys. In so doing, they utilize a wide range of metadata sources that include the full spectrum of resources associated with a survey effort (e.g., codebooks, instrumentation, response rates, etc.) as well as information about established procedures and processes for performing quality research in general and within the organization. They also often use administrative metadata (such

[7] This study was funded by an American Statistical Association/Bureau of Labor Statistics/National Science Foundation Research Fellowship and was only made possible by the access provided by BLS and the Census Bureau to documents and personnel. A full report of the study will be available in late fall at http://istweb.syr.edu/~cahert/asafellow.html

as hiring rules, project management rules), referral information (knowing who to inform or contact), and rationale information (why certain things were done, etc.).

The study participants indicated problems with accessing metadata they needed. The metadata that was most often lacking was rationale information—why something had been done a certain way but also what other options had been considered and the rationales behind their rejection. For example, a decision needed to be made about when to implement the new sample for the CPS (following the decennial Census updates). The year in which this sample is to be implemented is an election year and the participants in the particular meeting needed to 1) know what had been done the last time this had been the case 2) what the decision rationale had been and 3) what had been the pros and cons of the decision and the alternatives which had been rejected. The participants pieced some of this together from their collective memory but there was no indication that this information existed somewhere in a formal system.

*Decisions As A Fulcrum For Metadata Creation And Use*

Decision points were points at which the metadata creation and use process was clearly visible. Once a decision point is identified, decision makers might conduct research (thereby generating additional metadata) or by retrieving existing metadata. They may make a decision or table it. Several moments appear to be critical from the perspective of metadata retention. As decision makers move from identifying a decision point to acting on it, the question arises as to whether processes for examining existing information or on the appropriate decision criteria are developed or retained. The second point is the recording of rationale information discussed above. The third is when decisions are tabled. The Time Use Study team routinely generated issue lists (from minutes and other sources) that indicated outstanding decision areas and were referred to over time.

*The Context For Metadata Creation And Use*

An area of findings relates to the physical and organizational contexts in which the methodologists and metadata systems exist. Participants exploited aspects of the physical world as part of personal metadata systems (which connect to organizational systems) and the physical world shaped metadata usage and retention. The organizational context also influenced when metadata was necessary and how and where it was retained.

Physical aspects such as space and color are routinely used to provide context to enable understanding of information. Several respondents used different color files to convey meaning to themselves about the contents of the file (e.g., using different colors for different conditions in a cognitive test). Physical space (walls, desks, floors) is often used to position information to indicate relative importance or to facilitate ease of access to often needed information. Physical aspects are used to store metadata components perhaps obviating the need to record such information explicitly in text format.

Many respondents also talked about their retention of old files in conjunction with retaining the rationales behind decisions. But they also indicated that they rarely referred

to this files, weren't necessarily sure what was in them, or that they were not organized in useful ways. Extending the notion of retaining files, many people saved electronic mail messages that they deemed important for recording decisions. Most were less than satisfied in their ability to organize these or retrieve necessary information again.

The physical world also influences what metadata will get retained and in what formats. Many participants recorded information on "sticky notes", on white boards, etc. In some instances, the white board was used to communicate among the group. A more intriguing example is one in which metadata about a particular CPS case was not recorded electronically and the CPS staff in the regional office spent approximately a half hour trying to recall what the history was. All staff were included in the discussion, and it was pieced together that a case had been transferred to another field representative. This information had not been recorded, probably because the physical set-up of the office was such that all possible participants in the discussion were able to hear each other's conversations and "chime in". Such "chiming-in" was routinely done.

Finally, the organizational context is reflected in decisions about when metadata are created, where they are stored, etc. Obviously, organizational work practices, rules, and procedures often indicate what information must be retained, in what formats, etc. Additionally, the surveys (and metadata) under investigation here reflected aspects of the organizational environment in terms of their relationships to other agency data collection activities. Most obvious was the relationship of the CPS to the decennial Census activities (since every ten years sample had to be revised, etc.) and the ATUS to CPS used as a source of the sample and specific data on cases.

*The General Model*

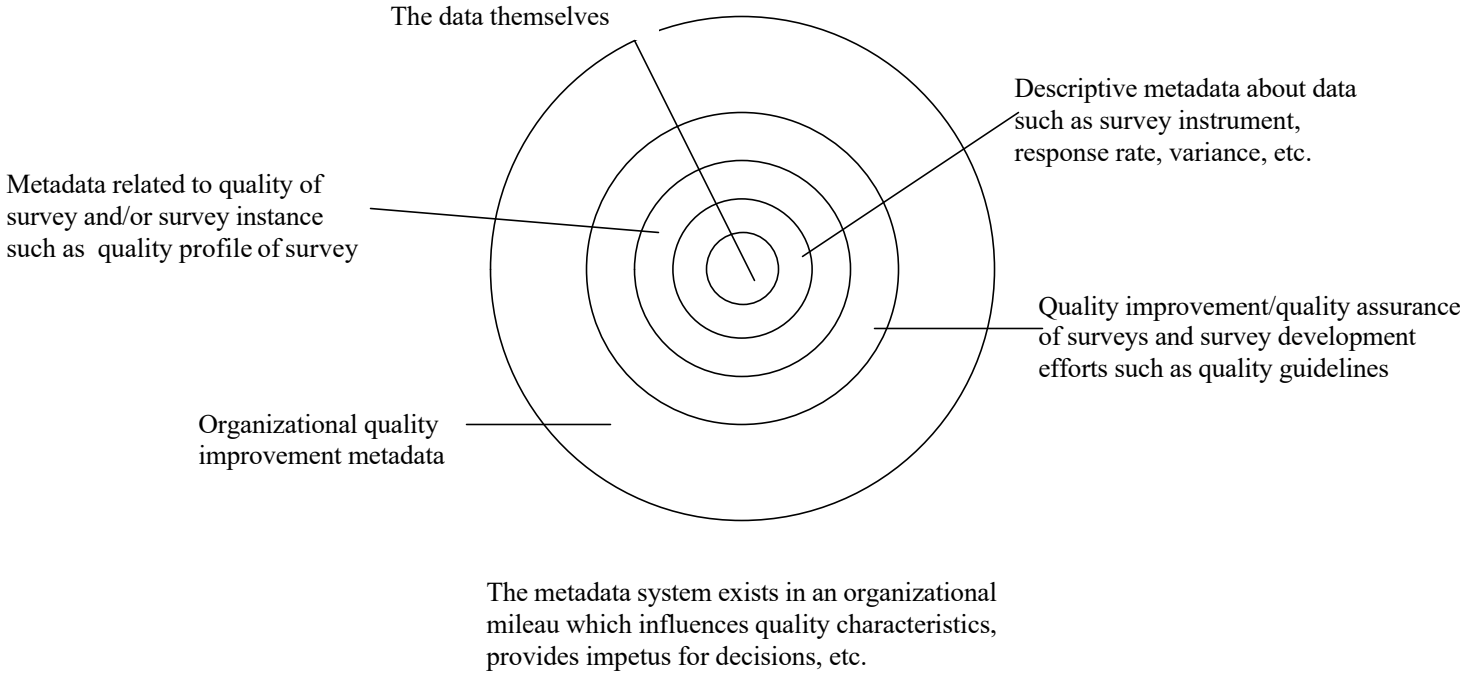
The following model (Figure 1) presents a quality-oriented perspective on metadata as a result of the analysis. Metadata created and used by survey methodologists might be conceptualized as a set of concentric circles. The innermost unit around which all metadata revolves is the data. The nearest circle to the data contains metadata that describes the data (such as response rates, dates of survey, instrument, etc.) The agencies tend to have agreement on what components are necessary here. While the components are shaped by quality concerns at this level they do not depict quality assessments (e.g., while a survey might be described as having a 85 percent response rate, the assessment of the quality of that response rate is not embedded within that description-it is added by observer knowledge or other information). The next further layer includes quality assessments of the survey or a survey instance. Quality profiles of surveys are one common representation of the type of metadata associated with this survey. Beyond that layer comes layers associated with quality improvement of a survey (such as quality guidelines about conducting surveys) and with processes and information associated with quality improvement at the personal and organizational level.

All layers might be accessed by an individual. The ability of an individual and an organization to perform quality work (and record that quality) is reflected in the final quality of the estimates that result from survey processes. This system of layers is

embedded in an environment that influences what is considered quality and appropriate processes to attain it, that serves as an impetus for decisions to be made, which results in an individual or group accessing the layers of metadata. That is the "world" to which the metadata system needs to be responsive.

We might imagine such a model for any of a number of communities. In this case, the study focused on survey methodologists in two surveys and the example metadata indicated above reflects their needs. The field staff might have different metadata necessary to support their work at these layers.

Figure 1: Quality Oriented Perspective on Metadata




**Meta-Themes**

What have we learned from these three studies of metadata usage? While it would be a mistake to try to directly compare them given the disparate populations, methods and analysis techniques, some commonalities do exist.

*Context Matters*

The larger world in which a number, set of numbers, or a metadata item sits is important in a variety of tasks. In the FERRET metadata relevance task, respondents wanted the ability to compare metadata for variables as a guide to determining which variable was most appropriate. Having skip patterns available or, as one respondent preferred, and actual copy of the instrument enabled respondents to understand why a particular question was asked, and hence the potential for a variable to be relevant. For the novice

users, in the table study, and also for the survey methodologists, the rationales were an important part of providing meaning to metadata or data.

Rationales were indicated as important components of metadata systems in all three studies. Unfortunately, the tables and survey methodology investigations suggested that rationales are not well documented in the available metadata systems. Since need for rationales and the content of a particular rationale are highly contextual, highly dependent on an individual user's knowledge, it is not surprising that they are often not easy to retain or access. The document analysis in the survey methodology study indicates that rationale information does exist in many cases, but is largely in unstructured text formats within documents such as minutes, etc. While it is not unfeasible to capture and retain rationale information appropriate to any user's need, analysis of particular populations can yield common needs for rationales. The survey methodologists tended to assess decisions (and thus rationales) using a constrained set of criteria (e.g., cost-effectiveness, quality, etc.) and it does seem possible to build a rationale metadata component that could address those criteria.

A third component of context is that of the physical world. Observing people at work, whether in the constrained FERRET or tables study or the more open-ended survey methodology study, made it quite clear that the physical world plays an important role in any metadata task. The studies didn't intend to document this specifically so no specific recommendations have resulted but metadata systems developers need to look at how people work in their worlds to understand preferences for access and storage.

*Metadata and Quality Assurance*

One valuable organizing principle for understanding metadata usage is quality assurance, particularly in the context of agency activities. Users external to the agency may have other critical dimensions in addition. For the agencies, however, the mission and goals revolve around quality data collection, analysis and dissemination and agency and individual activities can all be framed from that perspective. Thus we might consider metadata and metadata systems as knowledge repositories supporting quality efforts. The model presented as Figure 1 obviously needs validation. The next research phase might be to explicitly connect activities at various levels to data quality to assess the model's utility.

*The Value Of User-Centered Approaches*

Finally, the studies demonstrate that user-centered approaches provide rich and useful input in this environment. Since metadata are intended to be useful to people engaged in tasks, and because metadata systems are developed in a socio-technical context, an understanding of how real people interact with them provide signals to designers. For example, designers of metadata systems need to recognize that individuals have a need to have information close at hand for some tasks (thus, post-it notes or papers on desks) and a metadata system requires a person to log-on and execute a search might not be successful.

The studies also were able to identify places in which the metadata and metadata systems failed for a user or group of users. In the CPS FERRET study, users wanted to see metadata on multiple variables simultaneously and needed additional detail on universe statements. The survey methodologist study found that rationale information was generally not sufficiently retained or not easily accessible. The tables project pointed to specific metadata needs for a given table at a given moment in time.

This is not to suggest that metadata system designers should resolve all these limitations. These studies do not provide results that indicate the extent of a given problem—is the metadata element wanted one that everyone would use, or just a few people? The studies point to the need to further investigate and enable a researcher to design more targeted projects.

*Acknowledgements*

The research reported in this paper was done in collaboration with many individuals including Cathryn Dippo, Dan Gillman, John Bosley, and Fred Conrad at the Bureau of Labor Statistics; Elizabeth Liddy, Naybell Hernandez and Kristen Armstrong at Syracuse University, and Gary Marchionini at the University of North Carolina and Ben Shneiderman at the University of Maryland.

*References*

Dippo, C. and Sundgren, B. (2000). The role of metadata in statistics. 2nd International Conference on Establishment Surveys. Buffalo NY, June 2000.

Glaser, B.G. and Strauss, A.L. (1967). The Discovery of Grounded Theory. Chicago: Aldine.

Guba, E.G. and Lincoln, Y.S. (1989). Fourth Generation Evaluation. Newbury Park, CA: Sage Publications.

Hert, C.A. (1999). Federal Statistical Website Users And Their Tasks: Investigations Of Avenues To Facilitate Access: Final Report to the United States Bureau of Labor Statistics. Available at: http://istweb.syr.edu/~hert/BLSphase3.PDF

Hert, C.A., Hernandez, N.; and Armstrong, K. (in preparation). User Uncertainties and Their Relationship to Metadata in Statistical Tables. Draft paper to be made available at http://istweb.syr.edu/~tables/uncertaintiespaper.html.

Krippendorff, K. (1980). Content Analysis. An Introduction to Its Methodology. Newbury Park, CA: Sage Publications, 1980

Strauss, Anselm & Corbin, Juliet. (1990). Basics of Qualitative Research: Grounded Theory Procedures and Techniques. Newbury Park, CA: Sage Publications.