STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Joint UNECE/Eurostat Work Session
on Statistical Metadata
(6 - 8 March 2002, Luxembourg)

COMMISSION OF THE
EUROPEAN COMMUNITIES

EUROSTAT

Working Paper No. 15
English only

Topic (ii): Users and metadata, statistical information portals

CREATION AND USE OF METADATA IN TWO BUREAU OF LABOR STATISTICS SURVEY EFFORTS: AN ETHNOGRAPHIC INVESTIGATION OF A COMMUNITY OF PRACTICE

Submitted by the Bureau of Labor Statistics and Syracuse University, United States of America[1]

Invited paper

Abstract: The important role of metadata in survey development 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. This paper reports on an ethnographically conducted study of survey methodologists and associates, and their creation and use of metadata within the context of two survey efforts.

Findings are reported in four areas: 1) the expression of survey methodologists' work worlds as related to metadata processes, 2) metadata and their relationship to decision processes, 3) the context (physical, organizational) for metadata creation and use and 4) a general model for understanding the role of metadata in a statistical agency.

The study's findings have implications for agency practice. It identified additional types of metadata that agencies may consider retaining. The study demonstrated that this community used a wide variety of metadata in support of their work and that some types of metadata (most notable rationales) were difficult to access when needed.

A model of metadata types was developed that can inform discussions about metadata and that can be used to identify specific metadata that may be necessary within a survey effort. One of the challenges in creating metadata systems is the confusion about what is statistical metadata and what is not. The model developed in the study provides a strategy for different perspectives on metadata to be brought together. The study demonstrated the importance of understanding work practice in relationship to metadata system design. Efforts to build metadata repositories will not be successful unless they work for real people in real settings. The study provided both an approach to understanding those real worlds as well as specific insights into what aspects of the physical and organizational worlds of the respondents influenced their use and creation of metadata.

[1] Prepared by Carol A. Hert.

I. PROJECT CONCEPTUALIZATION

A. Defining Metadata

1. Metadata is an often ambiguous and nebulous term and is used variously in different communities. Dempsey and Heery (1998) (and Dippo and Gilman (1999) for statistical information) define metadata as information that enables one to manage and use the data/information to which they refer. Dippo and Sundgren (2000) further define statistical metadata as:

- ?? 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).

2. This definition highlights two key points, that metadata are defined within a context (there is no one set of metadata associated with a set of data), and that they are information that supports usage. Some of the purposes which metadata may support are information/resource discovery, administrative uses such as tracking terms and conditions of use, the context of creation, and unique identification of objects (see Bearman (1996) for a discussion).

3. Within the statistical domain, metadata may include subject heading schemes to support resource discovery (such as the list of headings employed by the American Statistical Index (published by the Congressional Information Service) and HASSAT[2] (from the University of Essex), codebook information, survey instruments and related documentation, as well as reports and other documentation produced by survey methodologists about data collection strategies, analysis of past survey efforts, etc. (Dippo and Gilman, 1999).

4. In the domain of the statistical agencies, one can identify a narrow definition of metadata that focuses on information about the data or specific data points (e.g., variances, response rates, response categories) and information used to produce the data (such as data collection instruments, instructions, technical documentation). In the study reported here, it was this definition that study participants employed when asked directly about metadata. There is also a broader sense of metadata that considers as metadata additional 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.).

5. 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 a broad range of information, is used to support a variety of tasks 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.

6. A variety of typologies of metadata exist. Colledge and Boyko (2000) frame metadata in terms of usage throughout the survey lifecycle and identify the following classes of metadata:

- ?? Definitional: Describing statistical units, populations, classifications, elements, questions, collection instruments, terminology;
- ?? Procedural: describing procedures by which data are collected and processed;
- ?? Operational: summarizing the results of implementing the procedures, including measures of response burden, edit failure rates, and other process quality indicators;
- ?? Systems: relating to data storage and retrieval;

[2] Available at http://biron.essex.ac.uk.services.zhasset.html

- ?? Dataset: a particular type of systems metadata required to describe, access, update and disseminate datasets including subject access, data cell annotations, etc.

7. Musgrave and Ryssevik (2000) take a dissemination perspective and categorize metadata as:

- ?? Catalogue (descriptive, primarily for discovery purposes);
- ?? Dictionary (Codebooks);
- ?? Context (explanatory information such as questionnaires, teaching materials, etc.);
- ?? Quality;
- ?? People (who can help with explanation and interpretation);
- ?? Supervening (information generated by the user such as how data were used, tips, corrections).

8. The Organization of Economic Co-operation and Development (OECD) (no date) provides the following list of metadata items used to describe compilations of statistics:

- ?? Source;
- ?? Concepts and coverage;
- ?? Standards;
- ?? Data collection;
- ?? Data manipulation;
- ?? Data quality and timeliness;
- ?? Data transmission;
- ?? Data storage and manipulation by the OECD;
- ?? Output preparation and delivery by the OECD.

9. Each of these typologies was designed for a particular context and use. It is difficult to integrate the typologies given these differences. However, they provide insight into what is commonly included as statistical metadata.

B. Past Empirical Work on Metadata Creation and Use

10. The study of user interaction with metadata is not completely unknown. Within the traditional library and information science domain, there is a thread of research most commonly known as relevance judgment research that investigates how users make judgments on the relevance (variously defined and operationalized) or potential relevance of information units. Traditionally those information units have been articles and books, and users examine representations of those units (such as citations, which represent the metadata in this case) and indicate those they consider relevant or non-relevant. Users are asked about the criteria they are using in the judgments and how they make those judgments. The intent of this line of work has been to understand the phenomenon of relevance judgment, provide typologies of relevance criteria, and in some cases to suggest enhancements to the representations of the information units (See for example, Park (1993) and Barry (1994).) For example, if users indicate that having information on the chapter titles in a book is helpful, it may be suggested that such information be added to the description of the book.

11. The vast majority of work of this type has looked at books (using information on records in online library catalogues) or articles (using periodical databases with or without abstracts). Users may be asked to examine different representations of the same item such as a citation, a citation with an abstract, or the item itself. Only recently have other types of information entities such as maps (Gluck, 1996) and meteorological data (Schamber, 1991) been considered.

12. In the domain of statistical information seeking, the author and Bosley (as reported in Hert, 1999) have been investigating how experts and other users employ metadata within codebooks (in this case, from the Current Population Survey) as they choose variables for analysis. He and Gey (1996) allude to the value of the codebook data in choosing variables in a paper that discusses a system that might facilitate browsing of such data. This work has occurred in tandem with efforts by the statistical agencies

to standardize and harmonize metadata across surveys. Reports by Dippo and Gilman (1999) and Gilman et al. (1998) highlight some of these efforts[3].

13. What has emerged from this work is an increasing understanding that metadata usage is contextual in nature--the tasks in which people are engaged, what they understand about the domain, and other aspects of their current situation determine the utility of metadata. Additionally, a wide variety of information that metadata creators may not have conceptualized as metadata is also helpful to users. For example, in the metadata studies of Hert and Bosley, it has been found that information about the purpose of the survey and a glossary of terms might both be helpful; this information is not currently available in the metadata that were investigated in the work.

C. A User Perspective on Metadata Creation and Use

14. The work above has provided important insights into how we might understand metadata usage (within a particular context). However, the body of user studies to date has been limited in several respects: it has 1) focused primarily on bibliographic entities used within library contexts, 2) often investigated only one type of metadata (such as citations or codebooks) and 3) not investigated metadata usage *in situ.*

15. The last limitation is particularly critical at this juncture. Most organizations will have to make hard choices about what metadata to create and maintain. At present, adequate automatic means to extract necessary information to build metadata repositories are not available and thus we can expect significant investment in human capital to identify, formalize and store metadata. Given that investment, understanding how people work, their preferences for metadata systems, and other factors will be critical in developing cost-effective systems.

16. The situation above presents an argument for a research approach that begins with the users (or community of practice) and that seeks to understand how users define their realities and how they act within these realities. 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. This was the approach taken in this study; it took an ethnographic perspective employing interviews, observations of work, participant observation, and document analysis as data collection techniques and utilizing the grounded theory approach (Glaser and Strauss, 1967) to inductive analysis of data to generate findings. Such a strategy, while not yielding formally tested generalizations, does lead to a richness of understanding. Such richness provides the basis for further work in which variables are explicitly conceptualized and hypotheses and theories explored and tested.

17. The study reported on here 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, the definition of metadata (discussed in later sections) that was initially used in the study was extremely broad so that new types of metadata might be identified.

II. PROJECT OBJECTIVES

18. This project investigated *in situ* how metadata were created and used by a community of practice-survey methodologists within the Office of Survey Methods Research at BLS as well as colleagues at BLS and their counterparts at the Census Bureau[4] in the context of two specific survey efforts. Survey methodologists represent an important link in the chain of statistical information provision through their

[3] 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.

[4] As the Census Bureau administers the two surveys of interest to the study.

ongoing efforts to understand limitations in existing survey methodology and to develop mechanisms and procedures to improve that methodology. Investigating their use of metadata offered important insights into how to enhance existing metadata to support their efforts, identified additional metadata, and provided a rich description and model of metadata usage that could inform other metadata initiatives.

19. The specific goals of the project were to:

> i) Provide a detailed picture of metadata usage by the community of practice and model that usage in order to enable enhancement of existing metadata creation and implementation practices (for both the community and others).
> ii) Develop a set of recommendations related to metadata practices within BLS to support this community.
> iii) Add to the existing theoretical literature within statistics and information science on how metadata are used.

III. METHODOLOGY

A. Introduction

20. The activities of the study occurred between July 2000 and September 2001. The researcher was resident at the Bureau of Labor Statistics or at the Census Bureau for approximately 6 months (spread throughout the time period above). Since the study's approach was ethnographic and analysis was inductive, data collection and analysis occurred throughout the period. Specific decision points and related analytic activities are discussed in individual sections below following a general overview of the project's approach.

B. The Grounded Theory Approach

21. The approach to data collection and analysis was informed by the work of Glaser and Strauss (1967) who employ inductive strategies they term the "grounded theory" approach. The goal of the grounded theory approach is to develop a set of plausible generalizations (or theoretical ideas) describing the phenomenon of study, which is grounded in the data collected. The generalizations might be considered precursors to theories in that they tentatively suggest relationship associated with a phenomenon or suggest some plausible hypotheses to be empirically tested, but have not themselves been rigorously tested.

22. Early in the process, a researcher is involved primarily in collecting and grouping the data. The grouping process consists of assigning codes to data instances. The groups are called categories. Data instances may be grouped in multiple ways, since early in the analysis it will not be clear which categories will become fully developed and contribute to the final theory. Instances that are similar are grouped together as a category though it may not yet be clear to the research what properties characterize the set of instances.

23. As more data are collected, the emphasis shifts to understanding the relationships among data instances and categories of data instances. This enables the researcher to develop higher-level, more abstract representations of the data. These are often generated in the form of "hypotheses" which a researcher explores in the data, looking for disconfirming evidence, confirming evidence, and additional themes. This is an iterative process that terminates when "theoretical saturation" is reached. Theoretical saturation, as defined by Glaser and Strauss, is the situation where new data do not contribute significantly to the set of generalizations that have been developed and all the collected data can be expressed in the generalizations. Glaser and Strauss also provide the following description of when to terminate the process:

> When a researcher is convinced that his analytic framework forms a systematic substantive theory, that it is a reasonable accurate statement of the matters being studied, and that it is

> couched in a form that others going into the same field could use—then he can published his results with confidence. (Glaser and Strauss, 1967, p. 113).

Throughout the process, a researcher also seeks feedback from study participants and others as a strategy to confirm the accuracy of the generalizations being developed.

C. Data Collection Activities

24. The grounded theory approach does not prescribe appropriate data collection strategies. Instead, like other qualitative, ethnographic approaches to research, a researcher draws on a wide range of strategies including interviews, observational techniques, participation in the phenomenon, etc. as appropriate to the situation and phenomenon. This study employed interviews, document analysis, and some observation and participant observational techniques.

25. Data were collected via interviews with study participants, observations of meetings (and some work practice) in which study participants engaged, participant observation in meetings relating to statistical metadata, and through document analysis. Initially, the desire was to observe work practices *in situ* without extensive interviewing. This proved to be difficult as 1) people often worked spontaneously on a project, for example, if they met a person in the hallway and/or 2) worked in contemplative ways that would require questioning them as they worked which would either be intrusive or become an interview. The researcher was able to observe some work activities such as when a survey methodologist ran a training session on a new tool she was developing but these opportunities were limited.

Table 1: Summary Table of Data Collection Activities: July 2000-July 2001*

Number of participants	23
Interviews conducted Number of participants interviewed multiple times	45 15
Meetings observed	20
Participant observation sessions including meetings, workshop participation, etc.	10
Number of documents collected	Approximately 125**

* these numbers do not include informal discussions with BLS staff, nor time spent in “passive observation”

** estimated as some documents were viewed on website and/or were packets of many documents (such as a workshop proceedings)

Interviews

26. A total of 23 people were interviewed, most multiple times. Due to the inductive nature of the study, the interview protocols changed throughout the study. There were five specific periods of interviews:
Round 1: Interviews with Office of Survey Methodology Research (OSMR) staff with goal of attaining general background on BLS, surveys, and work of survey methodologists
Round 2: Interviews with survey personnel on current tasks (related to the two surveys), resources and information used during those tasks (some personnel were interviewed multiple times during this round)
Round 3: Interviews included elicitation of current tasks on surveys as well as focused on specific decisions and decision points; for several specific decisions, perspectives on the decision were gathered from all participants in study that were involved in the relevant survey.
Round 4: Final interviews to clarify existing knowledge gaps for researcher, final elicitation of current tasks.
Round 5: Debriefing following initial framing of study findings with four participants.

The researcher took detailed notes during these interviews and transcribed them.

Meeting Observations

27. When possible, the researcher attended meetings in which participants were engaged in work associated with surveys. Given the researcher's schedule and that of participants, no attempt was made at consistency of coverage of meetings. A total of 20 meetings were observed and ranged from regular monthly meetings of the CPS Steering Committee (consisting of both BLS and Census staff), other meetings with BLS and Census staff, and meetings in which survey team members engaged with other audiences (such as outside researchers, field supervisory staff). The researcher also attended two conferences during the study period at which she gathered background information relevant to the study. These conferences were the 2nd International Conference on Establishment Surveys (June 2000, Buffalo, NY) and the Federal Committee on Statistical Methodology Policy Meeting (November 2000).

28. The researcher did not participate in these meetings. She took detailed notes and collected any documents that were provided during the meeting.

Participant Observation

29. The researcher did participate in some meetings at which she was present in her research capacity. Many of these were meetings of an ad hoc group concerned with metadata issues. In addition, she presented a paper and participated in a UN ECE Working Group on Statistic al Metadata meeting held in Washington DC during Nov. 2000.

Document Collection

30. Throughout the study, documents were gathered for analysis. These documents included materials provided during meetings, known metadata sources (such as Technical Paper 63 for the CPS), and materials used by respondents during tasks. Documents were logged and used in two ways. They were important in "fleshing out" comments made by participants, particularly as they represented the history that was recorded either officially or unofficially. Some documents were also analyzed in detail to identify "proto-metadata" or sources of information that might later represent more formal metadata.

31. The documents that were analyzed in detail consisted of a set of approximately ten minutes of meetings generated by the Time Use Survey team. These were read closely, and specific types of metadata mentioned were coded, as well as instances that represented information that could form the basis of metadata in the future. For example, minutes from September 7, 2000 raised the issue of which households would be included and the answer "only civilian households with members aged 15 and older will be eligible" was provided. While not encoded within the minutes as metadata, this information would form the basis for universe metadata.

D. The Analysis: Key Decision Points

32. The inductive research process is characterized by fluidity of data collection and analysis processes but can also be expressed in terms of phases where the focus is on a particular activity (such as data collection) and where a phase is concluded when the confluence of data and their analysis suggests particular paths and particular hypotheses to explore. In this study, these phases were

i) Preliminary data collection to understand survey methodologists' activities, metadata, and identify surveys to study further. The phase concluded with definition of a community of practice to study, particular surveys to investigate and preliminary sense of metadata.
ii) Interviews and other activities to further understand the specific surveys, the tasks individuals engaged in, and the resources they employed. The phase concluded with a refined definition of metadata, some preliminary hypotheses about metadata creation and use and their relationship to tasks and context.
iii) During phase three, data collection activities reflected the need to understand the hypotheses generated in phase two and their legitimacy. It ended with the integration of the hypotheses into

generalizations about the work worlds of the community of practice, the role of context, and a model reflecting the relationships among types of metadata.

E. Phase One: Framing the Boundaries of the Study

Communities of Practice and the Surveys

33. The users of metadata investigated in this study were members of a specific community of practice. A community of practice is defined as a group of people who share similar goals and interests and who have a common sense of purpose (Brown and Duguid, 2000). In the work context they are "informally bound to one another through exposure to a common class of problems [and] a common set of solutions" (Johnson-Lenz and Johnson-Lenz (undated website, http://awakentech.com, accessed 10/22/01). Common problems and solutions suggest common metadata needs, thus the focus on a particular community.

34. The operationalization of the community of practice was developed as part of the study. The original definition was the population of survey methodologists in the Office of Survey Methodology Research (OSMR) at the Bureau of Labor Statistics. This group consists primarily of Ph.D.-trained statisticians and cognitive psychologists. Initial interviews with members of this group resulted in a large array of projects in which they were engaged. In order to study their work, it would be necessary to understand each project at some level of detail; this was not possible within the time frame of the project. Therefore, the researcher, in consultation with members of the group, made the decision to focus on two specific survey efforts (discussed below).

35. The limiting of project scope in terms of specific survey efforts then led to a reframing of the operationalization of the community. Not all OSMR survey methodologists were engaged in tasks related to the two surveys and other personnel within BLS and at the Census Bureau were. Thus, the community of practice was redefined as personnel working on the two survey efforts who interacted with and worked on problems with the survey methodologists. This included economists, program directors, and program analysts. There was also some inclusion of field staff (the personnel with responsibility for executing the survey in the field) in order to understand the surveys more fully and to explore the boundaries among different communities of practice.

The Surveys

36. As indicated earlier, two surveys were the framing boundaries of the study of metadata creation and use. Preliminary interviews with survey methodologists in OSMR led to the identification of several dimensions of survey efforts that might influence metadata use. These dimensions were elicited by asking directly about metadata use and what might distinguish surveys in that regard. The dimensions were:

?? Extent of available metadata
?? Household survey vs. establishment survey
?? New survey effort vs. established survey
?? Researcher ability to understand purpose of survey
?? Traditional estimation methods used in survey or one using model estimation methods (researcher memo, 7-17-2000)

37. With these dimensions and with specific suggestions of surveys from the OSMR staff, the Current Population Survey (CPS) and the American Time Use Survey (ATUS) were chosen for further investigation. The two studies represent diversity in extent of metadata and length of time in existence. They are similar in that they are both household surveys and use traditional estimation methods. The researcher had previous experience with CPS and the ATUS was intuitively understandable to her.

38. CPS has been in existence for over fifty years and is "the primary source of information on labor force characteristics of the United States population" (http://www.bls.census.gov/cps/overmain.htm). It is a well-established survey and has a substantive and formalized set of metadata (much of which is

available at http://www.bls.census.gov/cps/mdocmain.htm. As a well-established survey, many of its procedures and processes have been made routine and there is an established organizational (and inter-organizational) structure. The American Time Use Survey is still under development and it is anticipated that data will first be collected in 2003. Its purpose is to measure how Americans spend their time, particularly in areas such as work without pay and child and elder care. Unlike the CPS then, it is in the process of creating metadata, determining workflow, and a variety of other aspects of the survey. The ATUS and CPS do share some connections; both are BLS surveys, fielded by Census and additionally, the sample frame from which ATUS sample will be drawn is retired CPS sample.

39. The activities of the surveys were investigated over a period of approximately one year (though the data collection activities were concentrated in the first 9 months). This did not enable a "full" view of the surveys as some processes occur in much longer time frames. For example, within the ATUS, the investigation was limited to a small part of a survey design process (in this case, related to early questionnaire design, and early field tests). CPS, while on a monthly data collection cycle also has other cycles such as the ten-year cycle associated with the Decennial Census. Thus this study might be considered to provide a "snapshot" of these surveys rather than a full longitudinal investigation. We might consider the units of analysis of the study to be the processes and activities engaged in by this community of practice during this particular time period on these two particular surveys.

Defining Metadata for the Project

40. In the first round of interviews, it became clear that the term "metadata" conveyed certain meanings for participants. Staff often introduced the term themselves based on their preliminary knowledge of the research project. This raised the concern that new types of metadata might not be considered or discussed because people had a preconceived definition they were using. Thus, the study began with an extremely broad definition of metadata: metadata is the information/knowledge that provides context for the task at hand.

41. With this definition, early interviews of phase two focused on understanding respondent tasks related to the two surveys (within the week the interview occurred) and what information or knowledge was brought to bear on those tasks. The sources of this knowledge or information were also identified.

42. As the project continued, the definition of metadata began to be constrained. Reactions from various stakeholders experienced with statistical metadata (for example, the UN ECE working group) about the breadth of the definition (it could include anything) and the researcher's recognition that such a broad definition did not provide boundaries or indications of what was metadata and what wasn't, led to a changed definition. The final definition used in the study was:

?? Metadata is information that performs the task of providing context designed to help the user of the metadata locate, understand, and use the entities/data to which the metadata refers.

?? Metadata is information preserved in some artifact (thus information in a person's head would not be metadata nor would verbally communicated information that is not recorded.

F. Phase Two: A Focus on Tasks

43. The activities of phase two were centered around data collection activities to gain a rich understanding of the tasks of respondents and how they used information and resources (which represented the broad definition of metadata) during the tasks. During this period, individual interviews were conducted and the researcher asked respondents to report on their activities during the current week and what resources they had used in support of those. A week was considered long enough to get a sufficiently rich picture but short enough for respondents to be able to report on their resource use. In addition, three and a half days were spent offsite at a Regional Office observing activities during several weeks of the CPS cycle and included some travel with a field representative.

44. This period was characterized by a growing set of "emphases," or ways to make sense of the data, that needed exploration. A researcher memo dated July 25, 2001 indicates an early set: cascading

sets of metadata (later to become the layers of the final model), the "locality" of metadata (how people used physical space), ownership of metadata, and processes by which people transform metadata into metadata they can use. A somewhat later memo (August 11, 2001) includes the following: types of metadata used, aspects of metadata in context, and possible models of metadata in context. Throughout the period, the researcher was engaged in collecting more data to add to these emphases or find new ones that integrated more of the data.

45. As data collection continued, the competing emphases (themes) were abstracted (or distilled) into several major threads. These were:

?? Revising the definition of metadata. Up to this point, the definition of metadata had been extremely broad. At this point, the definition was constrained to that subset of information/knowledge/data that is preserved as an artifact and which performs the task of providing context for the entities to which it refers;

?? A set of metadata types used by the community of practice;

?? Metadata as a potential strategy to communicate across boundaries (such as that between field staff and the survey methodologists);

?? The social/physical context in which metadata were created and used;

?? Metadata and decision processes-how metadata are created and used in decisions.

46. Some of these threads were theoretically saturated at this point. No new data categories of metadata were being identified in interviews, nor were more examples of the physical context necessary to support the theme that people use their physical space to organize themselves and their work. Thus, the researcher did not pursue additional data to flesh out these themes further. The Phase three activities were instead focused on the other themes which still needed further understanding.

G. Phase Three: Developing the Final Generalizations

47. Data collection in phase three continued to consist of interviews and observations. The interviews, while still asking about tasks, also began to include questions on decisions made, resources used in those decisions, and criteria for making those decisions. The researcher asked respondents to highlight decision processes in which they were engaged and two were chosen for detailed analysis. These were two decision areas associated with the CPS—the use of new race and ethnicity questions and the integration of SCHIP cases into the CPS estimates. The researcher conducted interviews specifically to gather detail on the "lives" of these decisions.

48. At the same time as data collection continued, the researcher began to express final generalizations and explore her data for confirming and disconfirming evidence. The generalizations were also introduced to some study participants to gather feedback on their legitimacy. During this period, several rounds of interviews were conducted as the researcher drafted various statements of the findings. The final interviews were done in July 2001 with five participants who reviewed the findings as they are reported here (in less detail).

IV. FINDINGS

The Work Worlds Of Survey Methodologists

49. 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.

50. 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 as hiring rules, project management rules), referral information (knowing who to inform or contact), and rationale information (why certain things were done, etc.). Table 2 provides the list of metadata types identified in the project (using the definition of metadata adopted in the project).

Table 2: Metadata Types Identified During the Study

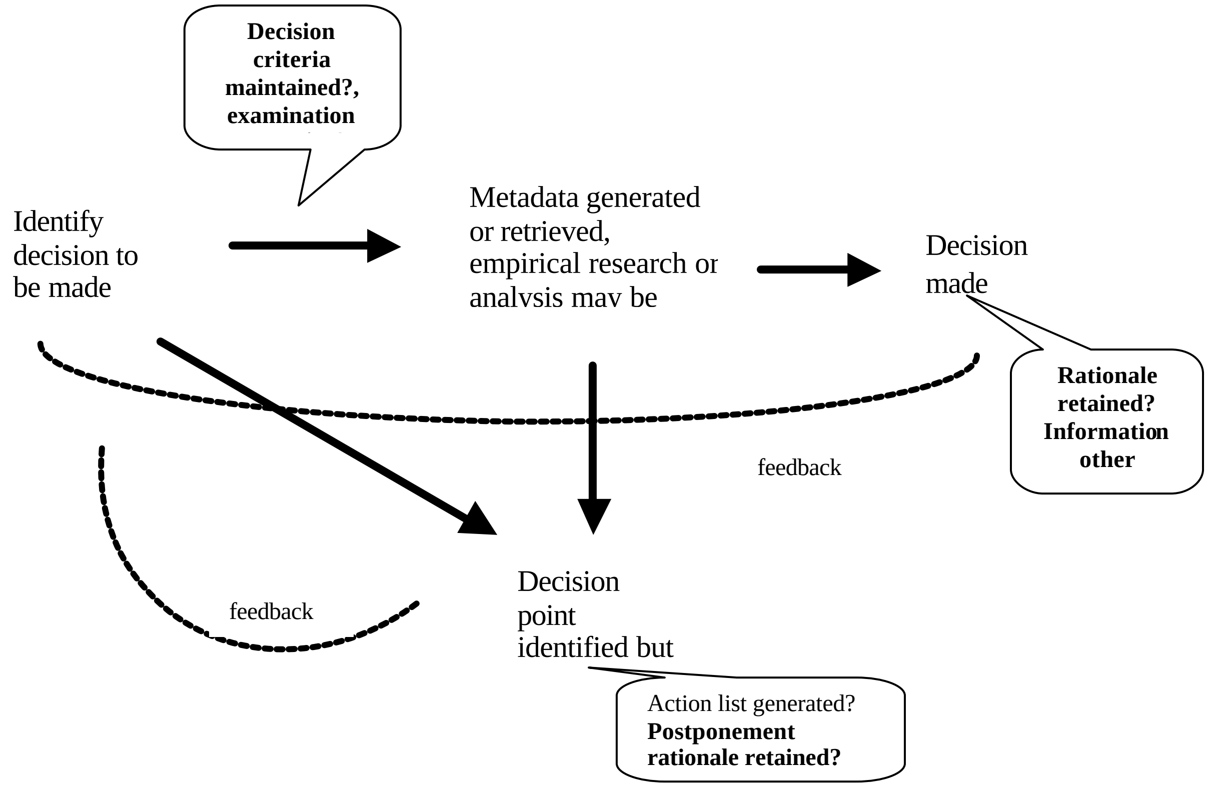
Metadata Type	Definition	Example(s)
Statistical	explicitly related to the final statistics including all categories mentioned by Musgrave and Ryssevik (2000)	Codebooks, interviewer instructions, response rates, etc.
Survey methodology	concerning survey design and implementation procedures	Textbooks about survey methods, rules/guidelines for conducting surveys
Agency-specific survey methodology	concerning specific policies, procedures for survey design and implementation within the agencies	Policies on appropriate formats for design specifications
Research methods	Related to non-survey methodologies	Methodologies for conducting cognitive tests
Project Management	Related to documenting projects and keeping them on schedule	Task lists, flow charts
Administrative	Related to general operations of the agency	Information about staffing rules, budgeting procedures
Referral/person	Information that enables someone to get additional help from a person	Telephone lists of contacts
Rationales	Explanations of events/actions	Explanation of a given universe choice

51. 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. Another example is this question from a August 2000 email message exchanged among CPS staff: “Does anyone know why we have a flag on the CPS Unit Control File (CUF) to indicate whether a CPS sample unit received the census long form? … There is a question of whether we need it for 2000 and without knowing what it’s used for, it’s hard to say yes or no.”

Decisions As A Fulcrum For Metadata Creation And Use

52. Decision points were points at which the metadata creation and use process was visible. Several key points in a decision process can be identified. There are 1) determining a decision needs to be made, 2) gathering knowledge to assess outcomes 3) acting on a decision (and recording that decision) or tabling the decision. Figure 1 represents a model of these process with "metadata junctures" identified.

Figure 1: Metadata creation and use during decision processes


Decision criteria maintained?, examination
Identify decision to be made
Metadata generated or retrieved, empirical research or analysis may be
Decision made
Rationale retained? Information other
feedback
feedback
Decision point identified but
Action list generated? Postponement rationale retained?

53. Once a decision point is identified, decision makers might conduct research (thereby generating additional metadata) or retrieve existing metadata. At this point, decision criteria will come into play. These criteria indicate what metadata to retrieve. Thus, one respondent indicated that his decision criteria included cost, effects on response rate, employee satisfaction, efficiency, and continuity with established organizational culture and procedures. Many respondents indicated that quality would be the overarching criterion. In the instance of decisions about integrating SCHIP cases into the CPS estimates, criteria such as stakeholder buy-in (from the states), and political consequences (including impact on the annual ASU exercise in LAUS) were mentioned. To make the decision resources in some of these areas would be gathered to assess the possible outcomes. Respondents relied on both existing metadata, their own personal knowledge of the organization and past actions, and also generated additional metadata through research processes or analyses of existing metadata.

54. 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.

55. After the preliminary identification of decisions and decision paths as important, two specific decision sequences were investigated in detail. These were the decision to integrate data from the SCHIP sample into the regular CPS estimates and the consideration of new Race and Ethnicity questions for the CPS. (Further information on these case studies will be available in the next draft of the report.)

The Context For Metadata Creation And Use

56. 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.

57. 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.

58. 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.

59. 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, perhaps 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.

60. 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 which will be used as a source of the sample and specific data on cases.

The General Model

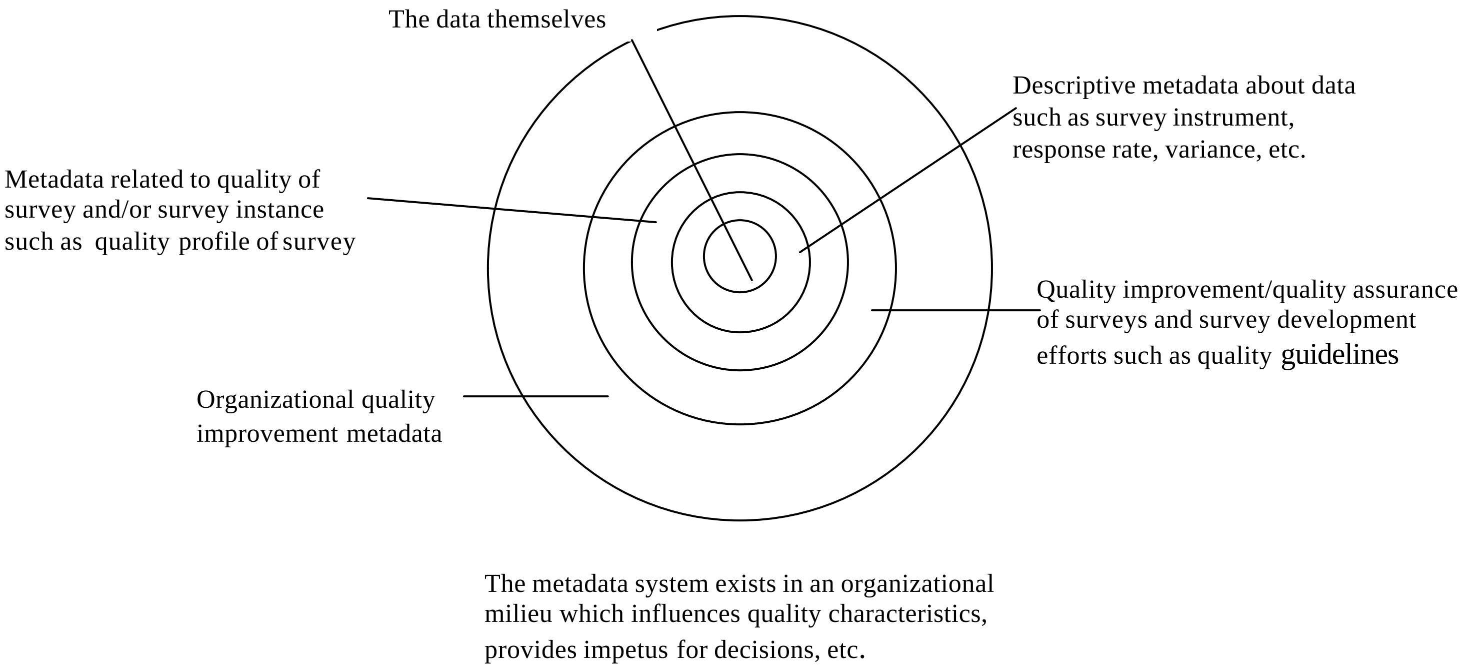
61. The following model (Figure 2) presents a quality-oriented perspective on metadata as a result of the analysis. Metadata created and used by survey methodologists might be conceptualized as a set of concentric circles. The innermost unit around which all metadata revolves is the data. The nearest 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.

62. An individual in the course of his or her activities might access all layers. 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. For example, survey methodologists have a concern with factors that impact response rate. To understand why a particular survey instance had a given response rate, they might first access the response rate (inner-most level), investigate that response rate by comparing to other survey response rates (quality assessment levels), and also gather information on pay scales of field staff, specific policies for allocating cases within the field, etc. (The outermost layer).

63. 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, and that is both physically and organizationally constrained. This environment results in an individual or group accessing the layers of metadata. That is the "world" to which the metadata system needs to be responsive and supportive.

64. 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, for example, might have different metadata necessary to support their work at these layers.

Figure 2: Quality Oriented Perspective on Metadata


The data themselves
Descriptive metadata about data such as survey instrument, response rate, variance, etc.
Metadata related to quality of survey and/or survey instance such as quality profile of survey
Quality improvement/quality assurance of surveys and survey development efforts such as quality guidelines
Organizational quality improvement metadata
The metadata system exists in an organizational milieu which influences quality characteristics, provides impetus for decisions, etc.

A. Meta-Themes

65. Have we learned from this study of metadata creation and usage? Several specific theoretical and practical insights result.

Context Matters

66. The larger world in which a number, set of numbers, or a metadata item sits is important in a variety of tasks. In particular, rationales were indicated as important components of metadata systems. Unfortunately, the investigation 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, and 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 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 feasible 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.

67. A third component of context is that of the physical world. Observing people at work made it quite clear that the physical world plays an important role in any metadata task. The study did not 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.

68. A final component is that decision-making often happens in meetings where access to formalized metadata sources is limited.

Metadata and Quality Assurance

69. 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 2 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. This would also enable a better understanding of the dimensions of quality.

The Value Of User-Centered Approaches

70. The study demonstrates 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.

71. The study also was able to identify places in which the metadata and metadata systems failed for a user or group of users. 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.

Identifying Metadata As It Is Being Created

72. A critical issue in the study of metadata is understanding how to identify it while it is being created rather than after the fact (and discovering that it is not available. The decision model (depicted in Figure 1) provides a strategy for capturing metadata or "proto-metadata" as it is being created. Not all of it may be useful over time, but this study has demonstrated that rationale information and the history of past actions is often the most difficult metadata to retrieve thus capturing it at the point of decision-making may be useful.

73. At this point, it is not clear in what format to retain this information. A follow on study looking at existing documents and identifying instances of metadata in creation is underway which may provide insights into appropriate storage strategies.

74. A second area to consider is the connection between the individual creation of metadata and its utility at the organizational level. When does information generated for individual usage (or within a particular survey) become valuable to be retained organizationally?

Metadata to Communicate Across Boundaries

75. If each community of stakeholders has its own concentric circles of metadata (ala Figure 2) then it might be possible to mesh the models to indicate communication across boundaries. Quality is critical to all members of the survey efforts but how it needs to be expressed and considered may differ based on what the task is or the particular staff person's responsibility. The field staff are concerned about getting quality data efficiently and so staff allocation becomes an important dimension of the quality equation and may be less important to the survey methodologist back at central headquarters. But the underlying issue remains the same and the meshing of the metadata models may provide a vehicle to, as one respondent phrased it "distill the quality context and communicate it."

Metadata and Knowledge Management

76. This study has demonstrated the overlap of initiatives in metadata management and knowledge management. Operational aspects of the organization and personal work behaviors (represented in the outermost layers of figure 2) suggest their importance to the quality of the final data. Understanding how metadata systems support knowledge management initiatives and vice versa has the potential to add value to both sides.

References

Barry, C.L. (1994). User-defined relevance criteria: an Exploratory study. Journal of the American Society for Information Science, 45(3):149-159.

Bearman, D. (1996). Developments in metadata management frameworks. Archives and Museum Informatics 10(2):185-188.

Brown, J.S. and DuGuid, P. (2000). The Social Life of Information. Cambridge, MA: Harvard University Press.

Colledge, M. and Boyko, E. (2000). Collection and classification of statistical metadata: The Real world of implementation. International Conference on Establishment Surveys-II. Buffalo NY, June 17-21, 2000. (unpublished paper, available from authors).

Dempsey, L. and Heery, R. (1998). Metadata: A Current view of practice and agreements. Journal of Documentation 54(2):145-172.

Dippo, C.S. and Gilman, D.W. (1999). The Role of Metadata in Statistics. Working Paper UN/ECE Work Session on Statistical Metadata, Geneva, Switzerland, Feb. 1999.

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

Gillman, D.W.; Appel, M.V.; and Highsmith, S.N. (1998). Building a Statistical Metadata Repository at the U.S. Bureau of the Census. Working Paper UN/ECE Work Session on Statistical Metadata, Geneva, Switzerland, Feb. 1998.

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

Gluck, M. (1996). Exploring the relationship between user satisfaction and relevance in information systems. Information Processing and Management. 32(1):89-104.

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

He, J. & Gey, F. (1996) Online codebook browsing and conversational survey analysis. Social Science Computer Review 14(2): 181-186.

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

Johnson-Lenz, P. and Johnson-Lenz, T. (undated) Awakening Technology website. Available at http://awakentech.com, accessed 10/22/01)

Musgrave, S and Ryssevik, J (2000).Beyond Nesstar: Faster access to data. Paper presented at the 2000 IASSIST meeting. Available at http://www.faster-data.org/FASTER.doc (accessed 10/22/01)

The Organization of Economic Co-operation and Development (OECD) (no date) Statistics at the OECD, Main Economic Indicators – Metadata, List of Metadtat Items to Describe the Compilation of Statistics. Available at http://www1.oecd.org/std/mastnew.htm (Accessed 10/21/01)

Park, T. (1993). The Nature of relevance in information retrieval: An Empirical study. Library Quarterly, 63:318-351.

Schamber, L. (1991). Users' criteria for evaluation in a multimedia environment. ASIS Proceedings 1991, pp. 126-133.