THE ROLE OF METADATA IN STATISTICS

Cathryn S. Dippo, U. S. Bureau of Labor Statistics and Bo Sundgren, Statistics Sweden
Cathryn Dippo, Bureau of Labor Statistics, 2 Massachusetts Avenue NE, Rm. 4915, Washington, D.C. 20212
Dippo_C@bls.gov

ABSTRACT

Metadata plays a vital role in both the development and use of statistical information. The production of information requires that data and metadata be viewed as as a totality rather than individually; thus, metadata management must be seen as an integral part of statistics production. Moreover, since metadata provides the basis for human understanding of data, the cognitive aspects of metadata must also be addressed.

Key words: information, usability, users, dissemination, management.

The concept of ”metadata” and related concepts such as ”metainformation”, ”metadatabases”, and ”metainformation systems” were first defined in Sundgren (1973). A very short definition is that metadata is ”data about data”, that is, some kind of second-order data; cf Froeschl (1997). Among computer scientists the meaning of metadata is often limited to formal descriptions of how data are typed and formatted. Information scientists and system developers, on the other hand, also stress the importance of metadata as descriptions of the meaning or semantical contents of data. These descriptions may be more or less structured and more or less formal; they are often free-text descriptions.

Official statistics was probably the first area to recognize the importance of metadata, but even there it took about two decades (and a number of unsucessful projects) until some real progress could be seen. During the 1980’s and the 1990’s the Statistical Division of UN/ECE organized several meetings on statistical metainformation systems (METIS). One tangible result was a Guideline; Sundgren (1993). In 1993, Eurostat arranged a workshop on statistical metadata that attracted a lot of attention and a large number of participants. In 1994, the Compstat conference had a session on statistical metadata.

Only recently other sectors of society, including the private business sector, have felt the need for a more comprehensive and serious approach to metadata. To some extent these needs have been triggered by the interest of companies and organizations to reuse their operational data for more strategic purposes, by organizing the data in so-called data warehouses, and by using new techniques like On-Line Analytical Processing (OLAP) and data mining. Such secondary usage of data generated by an organization´s operational procedures obviously have a lot in common with production and usage of official statistics (which to a large extent rely on operational data generated by a society´s administrative system). In both cases metadata are essential for compensating for the distance in time and space between the source and the usage of the data; for example, a user of historical data may not even be born, when the data he or she is interested in were collected and stored.

Powerful tools like databases and the Internet have vastly increased communication and sharing of data among rapidly growing circles of users of many different categories. This development has highlighted the importance of metadata, since easily available data without appropriate metadata could sometimes be more harmful than beneficial. Which producer of data would like to take the risk that an innocent or malevolent user would, in the absence of appropriate metadata, inadvertently or quite consciously misinterpret data to fit his or her own purposes? Even if data are accompanied by complete, high-quality metadata, such misuse cannot be completely avoided, but if it occurs, there is at least an objective information basis to argue from.

Metadata descriptions go beyond the pure form and contents of data. Metadata are also used to describe administrative facts about data, like who created them, and when. Such metadata may facilitate efficient searching and locating of data. Other types of metadata describe the processes behind the data, how data were collected and processed, before they were communicated or stored in a database. An operational description of the data collection process behind the data (including e.g. questions asked to respondents) is often more useful than an abstract definition of the ”ideal” concept behind the data.

There are several examples of existing metadata standards. For example, the Dublin Core (see http://purl.org/metadata/dublin_core) is a set of 15 metadata elements intended to facilitate discovery of electronic

resources. Metadata content standards now exist for a variety of subjects, including biological and geospatial data (http://www.fgdc.gov/metadata/contstan.html).

It is a less complex task to develop general standards for formal, technically-oriented metadata than to do the same for less formal, contents-oriented metadata. Thus, most general standardization efforts concern the computer scientists´ concept of formal metadata, whereas contents-oriented standardization of metadata is more dependent on the particular context or universe of discourse of the data, and, hence, often takes place within specific application fields, such as biology, geography, or statistics.

But what does the term ”metadata” mean with respect to our field of official statistics? While the dictionary definition ”data about data” is concise and accurate, it lacks the specifics and context needed to communicate meaning. So, a few years ago, members of the Open Forum on Metadata developed the following definition: ”Statistical metadata describes or documents statistical data, i.e. microdata, macrodata, or other metadata. Statistical metadata facilitates sharing, querying, and understanding of statistical data over the lifetime of the data.” This definition is fairly concise and accurate; moreover, it provides some context. But is it sufficient to convey meaning to a diverse set of users such that their comprehension of the term is equivalent? Probably not.

To be more explicit in defining statistical metadata, one must discuss the fundamental role of metadata. Metadata provides context for data; without metadata, data has no meaning. Thinking mathematically, data coupled with metadata as a set yields information. For example, the number 4.1 is just a number until one is told that the number is the official estimate of the seasonally-adjusted unemployment rate in the United States for the month of May 2000 as published by the Bureau of Labor Statistics on June 3, 2000.

Depending on your intended use of the number 4.1 and your general knowledge, the metadata given above may or may not be sufficient. If you have a general knowledge of statistics and the concept of uncertainty, you may want to know an estimated confidence interval or coefficient of variation. If you are a policy analyst, you may want to know the detailed definitions used for classifying someone as employed, unemployed, or not in the labor force. If you are knowledgeable about survey methods, you may want to know the response rate or maybe even the form and sequence of questions used. And this is just a small beginning with respect to describing the metadata available for this one number.

Our goal in this paper is to indicate the breadth of meaning associated with the term metadata in the context of official statistics and the agencies that produce them. First, we examine the why, who, what, when, where, and how of statistical metadata. We show that a diversity of perspectives is needed to describe statistical metadata. In section 2, the relationship between metadata and quality are discussed. In the last two sections of this paper, we describe some of the multidisciplinary research efforts currently underway at the U. S. Bureau of Labor Statistics and at Statistics Sweden. The results of these projects will help us clarify the definition of statistical metadata across a wide diversity of users and usage.

1. DEFINING STATISTICAL METADATA: WHY? WHO? WHAT? WHEN? WHERE? HOW?

One lasting insight from many years of analyses, discussions, and experiments is that statistical metadata issues need to be treated in several dimensions: why? who? what? when? where? how? This is the topic of this section. Another important insight is that the metadata of an organisation have to be regarded as a system. Otherwise, it will not be possible to satisfy all the important needs for metadata with the time and resources available. This topic will be treated in section 4.

1.1 Why are statistical metadata needed?

Statistical metadata have several purposes. The first and most fundamental purpose is to help a human user of statistical data to interpret, understand, and analyze statistical data (microdata, macrodata, or other statistical metadata), even if they have not themselves participated in the production processes behind the statistical data. In other words, statistical metadata should help a human user to transform statistical data into information. (See Hand (1993) for an excellent discussion ”Data, metadata and information.”)

Information is only in the brains of people. Information can only be communicated and shared between people by means of data representations. Information can be represented by data in many different ways: spoken and written languages, pictures, electronic representations, gestures and body language, etc.

Statistical metadata also helps a user to identify, locate, and retrieve statistical data of possible relevance to the user's information need. Statistical information seeking, especially in this Internet age, is a task that has begun to receive some attention in the information science community (see section 3), but many of the problems that have been discovered have no easy solutions. One set of very important and persistent problems relates to concepts and terminology, i.e., the mismatch between producer´s and user´s concepts and the fact that technical terms can have multiple, contradictory definitions (even in a single organization). Metadata can help to solve such problems.

Statistical metadata, and, in particular, so-called process data are used to describe and provide feedback concerning all subprocesses and steps that occur in a statistics production chain, operation processes as well as design and planning processes. These metadata are indispensable for evaluators of statistical production processes, including the producers themselves. Most methods of process improvement, including those of Deming (1982), are built on the availability of metadata or data about the production process. The same kind of process descriptions may also be valuable for instructional and training purposes, e.g., when introducing new staff or improving the performance of existing staff.

Statistical metadata documents existing surveys, production systems, and production tools in such a way that these resources and experiences can be used by designers of new surveys and production systems. Thus, statistical metadata can be used in knowledge bases and knowledge-based systems (e.g., expert systems), and for knowledge management purposes, in general, in connection with the design and operation of statistical surveys and production systems. For example, consider how difficult it would be to develop a new survey questionnaire that is to provide information on the health care of children in poverty if one does not have access to the standard set of questions used to classify a family as being in poverty.

Statistical metadata describes statistical data in such a way that it can be processed by computer software. Such metadata need to be more structured and formalized than metadata intended for human users of statistical data.

Thus, the primary role of statistical metadata is one of facilitation and sharing. Metadata is necessary for the interpretation of statistics. The new knowledge gained from interpreting statistics may lead to production enhancements (lower costs or better quality) or the creation of intelligence or understanding about some real-world phenomenon. Moreover, metadata is data for the survey designer. Its compilation and storage aid the designers of new measurement processes through reuse or learning from the past.

1.2 Who uses statistical metadata?

There are two broad classes of statistical metadata users—the producers and the users of statistics. By producers, we mean the designers of data collection processes, data collectors, data processors, and data evaluators, i.e., everyone in the agency and its contractors that plays even a minor role in the development, production, and evaluation of statistics. The user group includes civil servants, politicians, policy analysts, social scientists, financial analysts, students and teachers at all levels, journalists, and interested citizens.

Different users have different requirements for statistical data and metadata. They also differ in resources and abilities. Thus, there are many different user profiles that we have to take into account when designing statistical metadata and statistical metadata systems.

Producers of statistics may also be users. However, there is an important distinction between an ”in-house user” and an external user of statistical data that should be taken into account when designing metadata and metadatasystems. A producer-user has meaningful relevant pre-knowledge thanks to his/her own participation in the design and operation of the statistical production processes. Thus, an in-house producer-user will not have the same need for metadata as an external user who has not participated in the design and production of the statistical data.

1.3 What is statistical metadata?

A simple, basic definition is that metadata are data that describe other data. Thus, statistical metadata are data that describe statistical data. Statistical metadata may also describe processes that collect, process, or produce statistical data; such metadata are also called process data. Finally, the term ”statistical metadata” may also be used for descriptions of resources and tools that are instrumental in statistics production, e.g., statistical classifications and standards, registers, and statistical methods, procedures, and software.

Since the metadata needs of users vary greatly, the definition of a necessary and sufficient set of metadata also varies by user and usage. For example, users looking for a number specified by a contract or lease only need a minimal set of metadata — enough to locate the specific number needed. On the other hand, the survey designer evaluating data quality from alternative data collection procedures requires a great deal of metadata. That is, if, for instance, respondents are given a choice in the mode of response (e.g., mail, touchtone, Internet), the evaluator needs to know the specifics of each mode (e.g., physical layout or type of voice, means of navigation) and how each respondent interacts with the particular mode they chose (e.g., missing item responses, backups or hang-ups). Since there is no detailed, causal model of nonsampling error, there is no way to specify the minimally sufficient set of metadata needed to evaluate alternative designs or quantify the quality of a specific design. Consequently, a designer's or evaluator’s view of metadata is constrained only by his ability to define what he thinks is relevant metadata.

Another example: A journalist will have neither the competence nor the patience to digest large volumes of detailed, theory-based metadata; instead it is urgent to provide such a user with powerful, pedagogically presented metadata that helps him or her avoid the worst interpretation mistakes. On the other hand, a social scientist may even want to question the assumptions made by the original producer of statistics and derive new statistical results on the basis of alternative assumptions. The latter kind of user will need to have access to all assumptions and other relevant circumstances in the data collection, data preparation, and estimation processes, as designed and operated by the statistics producer.

1.4 When is metadata used?

The production of statistical information is a complex process. No new data collection effort or revision of an existing one takes place in a vacuum. Metadata in the form of prior experience, whether recorded or from personal knowledge, is used by everyone involved in the creation and use of statistical information from the initial planning stages through the use of the products. The more relevant metadata is available to someone designing or implementing a particular procedure, the more likely the specification or result will be of better quality. The more metadata are linked to specific pieces of data or statistics, the more likely a seeker of information will find the appropriate number and make proper use of it now, tomorrow, or several centuries from now.

1.5 Where is metadata used?

The use of the word metadata, as opposed to documentation, is an important one. The word documentation has its semantic roots in a matter-based medium, primarily paper but also stone and metal (coins). Moreover, documentation is usually associated with writing. Metadata as part of statistical information is not confined to writing on paper. Maps, graphs, computer screen shots, computer programs, compiled code, scanned documents, and data bases are all components of metadata. Some only exist in cyberspace. Certainly, the use of metadata is not confined to buildings with four walls and a roof (e.g., offices, classrooms, homes); data collectors in the field collecting data on crops, water and air quality, fish and wildlife, etc. are heavy users of metadata. As we move towards a more digital environment in the production and use of statistical information, the places where metadata are used will only be limited by physical conditions that preclude the use of a computer.

1.6 How is metadata used?

Metadata is a tool for comprehension and understanding. It provides meaning for numbers. At the most basic level, metadata makes it possible to interpret a number. That is, the number 4.1 has no meaning without metadata. Metadata is also a tool for interpretation, using data to make inferences and facilitating the acquisition of new knowledge. Metadata helps the information seeker find data and determine if it is appropriate for the problem at

hand, i.e., determine its fitness for use. Metadata helps the designer develop new, improved processes and the implementer meet process specifications, e.g. by informing about relevant methods and tools, how they can be used, and what the experiences from earlier applications are.

Metadata is also a tool for modifying work processes to improve data quality or reduce costs. Documenting procedures with respect to what worked and what didn't will help others make better choices and avoid pitfalls. Reductions in costs can result from the reuse of metadata from a previous implementation (e.g., electronic data collection instruments, software for sample selection or weighting, a word processing document of an interviewer's manual).

1.7 Conclusion

In summary, the role of metadata is a ubiquitous one. Any and all definitions may be appropriate given the particular circumstances. So, how do we decide what is the appropriate set of metadata for a specific instance? Research. In the last two sections of this paper, we will describe recent and ongoing research projects designed to inform producers on the process of providing metadata to users. But first, a discussion of metadata and quality.

2. METADATA AND QUALITY

Metadata plays a key role in linking survey measurement and process quality improvement (Dippo 1997). There is a bidirectional relationship between metadata and quality. On the one hand, metadata describe the quality of statistics. On the other hand, metadata are themselves a quality component which improves the availability and accessibility of statistical data.

2.1 What characterizes good quality statistics?

First, good statistics should be relevant for the user's problem. This has to be judged by the user in a concrete usage situation. The same statistics may very well be relevant in one usage situation and more or less irrelevant in another usage situation. The problem of relevance is a difficult one in official statistics, since such statistics are produced for many users and usages over a long time period, so-called multi-purpose statistics. In order to enable many users, now and in the future, to judge the relevance of certain statistics in many different usage situations, a lot of metadata have to be provided about the meaning of the originally collected data (possibly from different sources) and about how these data were treated in the original production process.

Second, good statistics should be reasonably correct (accurate, precise), that is, they should be free from serious errors. As a minimum, the sources of errors should be known (and documented), and, when possible, the error sizes should be estimated. Enhancing metadata on accuracy and precision should be an integral part of the statistics producer's work program.

Third, good statistics should be timely and up-to-date. Good, managed metadata can facilitate reducing the time lag between design and implementation by reducing development time through reuse (e.g., software components, questions, and procedures). Moreover, by managing metadata as part of the production process, the timeliness and quality of dissemination products can be improved.

Fourth, good statistics should be well-defined to facilitate comparability with other statistics that are needed by the user in a certain usage situation, e.g., similar statistics concerning another region or a time period. Comparability can only be confirmed through accurate metadata. Thus, it is necessary to manage metadata on changing classification systems and geography and the links between the data and metadata. Otherwise, a user might misinterpret differences as a change in the phenomenon being measured rather than a difference in geographic coverage or classifier.
Fifth, good statistics should be available, easy to retrieve, interpret, and analyze. Good metadata facilitates resource discovery, especially via the Internet. Thus, metadata content standards like the Dublin Core and the Data Documentation Initiative (DDI) are essential. The DDI committee has produced what is known as a Document Type Definition (DTD) for "markup" of codebooks for microdata sets. The DTD employs the eXtensible Markup Language (XML), which is a dialect of a more general markup language, SGML. The DDI is already in use by major

international projects such as the European Networked Social Science Tools and Resources (NESSTAR). (See http://www.icpsr.umich.edu/DDI/intro.html.)

2.2 The role of process data in quality declarations

It is not as easy to declare the quality of statistical data as it is to declare the quality of a physical commodity, like a car. In the latter case, ordinal scales (say 1 to 5) are often used to indicate good/bad quality for a number of important ”features” of the commodity. In the case of statistical data, there are few absolute features, which can be evaluated in the same way for all users and usages, known and unknown. There are many more features, which have to be evaluated by the user, taking into account the particular usage at hand. In order to enable a user to make such evaluations in a particular usage situation, the producer of statistical data and metadata must provide rather detailed descriptions of the processes behind the data, for example:

- What questions were asked, and how were they asked?
- How were the answers to the questions checked for possible errors and mistakes?
- What rules were used for imputing and coding data?
- What were the discrepancies between the target concepts and the measured concepts?
- How was nonresponse handled?
- What estimation assumptions and estimation procedures were used?

As a consequence, the production of good quality statistical metadata requires a commitment from the statistics producer, a commitment that fits hand-in-hand with a commitment to produce good quality data.

3. RESEARCH ACTIVITIES AT THE BUREAU OF LABOR STATISTICS[1]: USER STUDIES

Research activities related to metadata at the Bureau of Labor Statistics are focused on users. Activities include user studies and knowledge organization by information scientists, cognitive studies by cognitive psychologists, and usability testing by human factors psychologists.

3.1 User studies

Knowing who your users are, what they want, and their expertise is vital to the design of a usable, useful website that has sufficient metadata to allow users to be satisfied customers. Over the last few years, Marchionini and Hert (1997) studied users of three official statistics websites: Bureau of Labor Statistics (BLS), the Current Population Survey (a joint Census-BLS venture), and FedStats (a joint venture of the 14 statistical agencies which are part of the Interagency Council on Statistical Policy). In the first year, their goals were to determine who used these sites, what types of tasks they brought to the sites, what strategies they used for finding statistical information, and to make recommendations for design improvement. They used a variety of methods in their investigations. Many of them are similar to the methods used by behavioral scientists in developing and testing questionnaires, i.e., interviews, focus groups, and content analysis. One result of their research was the development of a query-based taxonomy of user tasks. They also suggested interfaces could better meet the needs of users with diverse expertise and needs if they did not reflect the organization's program-orientation. Based on these results, Marchionini (1998) iteratively developed and tested designs based on four design principles: user-centered, alternative interfaces for different groups of users (rather than interfaces that adapt to individual users), information abundant, and spatial display.

Hert (1998) in her follow-up study of users through interviews with intermediaries found a number of metadata-related problems. Some examples are: lack of knowledge about how data were collected, lack of mathematical and statistical capabilities, and lack of understanding concerning the research process or nature of error. Historically, intermediaries have provided the knowledge needed to address these deficiencies; however, for dissemination via Internet, the website must provide the metadata-based services currently provided by intermediaries. Examples of such services are tutorials, scenarios, and context-based online help.

[1] John Bosley and Fred Conrad of the Bureau of Labor Statistics contributed to the preparation of this section of the paper.

3.2 Usability testing

Usability laboratory testing to evaluate the human computer interface should be an integral component of any system development effort. This extends to design of statistical websites and other statistical data bases. Usability testing of statistical websites typically consists of asking a group of test participants to carry out some data-related tasks, such as selecting and downloading one or more variables, by manipulating objects that appear on one or more interfaces accessible at the website under scrutiny. In early, informal tests of "trial" interfaces, the participants may simply explore the interface(s) and comment on how useful various features appear to be, how they like the overall arrangement of interface objects, and the extent to which the site structure makes sense to them. These evaluations are fed back to the web designers, who then refine their design and put it through another iteration of usability tests. As the design matures, participants may be given structured tasks (scenarios) to carry out so that performance data capable of analytic scrutiny may be collected, e.g., the average time that a group of users takes to complete a given scenario, the proportion of times users retrieve the target data.

Video cameras may be used to record the subject's face (and verbal comments) and their interaction with the keyboard and mouse, and the resulting tape is then integrated with a video out from the workstation display. Researchers may observe the live test or view tapes, often editing tapes to highlight significant design problems. Usually there is a debriefing session after the tasks are completed in which the test team can explore issues with the participants that the observational data did not resolve satisfactorily. For example, participants can be queried about unexplained interruptions of task performance that were observed, to get their subjective accounts about reasons for such occurrences

An alternative approach (that need not be carried out in the lab) is to examine the way the users think about the information that the site is intended to make available. One way to do this is to ask users to sort cards containing the names of possible web page topics into piles and by visually inspecting or cluster analyzing these piles to determine the degree to which users' conceptions of how the information is structured correspond to designers'.

Human factors researchers at BLS have conducted a number of usability tests on the BLS internet and intranet sites, the CPS site, and the prototype user-based interfaces designed by Marchionini (1999) as an alternative to the current BLS home page. These involve the use of metadata to the extent that they evaluate users' abilities to retrieve documents that describe actual data. However, they have as much or more to do with improving the structure of the web sites, so that users can more easily locate and retrieve numerical data. The structure of a web site and the design of web pages are types of metadata; they provide information about location and context of data.

3.3 Cognitive studies

Laboratory experiments involving think-aloud interviews and other cognitive research methods can and should be used to understand website users' strategies for information retrieval and comprehension of the terms being used. That is, is sufficient metadata being provided to aid the user in retrieving and understanding what is presented?

Hert conducted an experiment with four variants of an A-Z topic index. She found that the structure of existing organization tools and the terminology employed in these tools is highly problematic for users. Thus, she recommended the index be enhanced by adding multiple entries for a topic and the entries be in the language of the general public.

BLS and Census researchers have conducted some pilot research directed toward developing conventions for assigning short names to survey variables. Rules and guidelines for building a naming convention are provided in Part 5 of ISO 11179. That naming convention, however, was generated from a model of data that is not clearly grounded in research on how a broad spectrum of data users may interpret the names or their components. In pilot work, different semantic and grammatical rules were used to generate variant names, and a small group (N=15) of data users made numeric ratings of how well each variant captured the meaning of the corresponding question. Analysis of these preliminary results indicated the variations in naming semantics had little influence on comprehension. On the other hand, even this small test indicates that it may be more difficult to find <u>any</u> "good" short name for certain types of variables. Further research will focus on testing and refining the latter preliminary finding. This additional research will also be redesigned so that test participants will actively construct names for

variables, using procedures developed by lexicographers for building dictionaries, instead of merely reacting to variant names created by the research team. This approach was suggested by another information scientist who has been working with BLS, Stephanie Haas (1999) from UNC-Chapel Hill.

Another ongoing project is aimed at identifying the minimum amount of metadata that data users need in order to make accurate and confident decisions about the relevance of a particular survey variable to a planned analysis. Preparation for this study involved creating a set of plausible research scenarios capable of being carried out using data from a widely-used BLS/Census data set, the Current Population Survey (CPS). Then, a group of veteran CPS data users at BLS reached consensus on the subset of CPS variables that are the "best" to extract in order to perform an analysis that would satisfy the goal of each scenario. These expert users also nominated a larger set of similar-sounding but less suitable CPS variables for each scenario, in order to force study participants to pick the best variables from a list of competing data items. In the actual study, the amount of metadata made available to participants about the variable lists will be set at three levels—minimal, moderate, and rich. Participants' choices of "best" variables will be compared across these three levels, to see how much having more metadata available improves accuracy of choice relative to experts' judgments. Participants will also provide data on which metadata elements they found most useful in discriminating between the most relevant variables and less suitable competing data choices. This line of research will continue with additional studies, to see if a "point of diminishing returns" for metadata can be roughly established, beyond which additional information does not improve users' choices among competing variables.

3.4 Conclusion

As noted in section 1.1, the first and foremost purpose of metadata is to help a human user of statistical data. For a statistics' producer to determine if it is providing usable, useful, and sufficient metadata, it must engage in user studies. The cognitive aspects of metadata and, for that matter, most components of statistical dissemination products (e.g., text, tables, charts, graphs, maps) is an area that deserves significantly more attention from statistics producers.

4. RESEARCH ACTIVITIES AT STATISTICS SWEDEN: INTEGRATED METADATA MANAGEMENT

It is quite obvious that statistical metadata have many different and very important users and usages. There is no doubt of the need and demand for statistical metadata. The supply side is more problematic. Who is going to provide the urgently needed metadata? The ultimate provider of statistical metadata can be no one else but the producer of the statistical data to be described. However, producers of statistics are not always well motivated to produce metadata as well. First of all, they (often wrongly) assume that they themselves know everything worth knowing about the statistics they produce. They carry this knowledge with them in their brains, and they see little reason why they should document this knowledge so that it could be shared by others in other places or at later times. "If somebody wants to know something about these statistics, they are welcome to ask me" is a rather common statement by producers of statistics. Such a comment disregards the fact that even a producer of statistics has an imperfect memory and that he or she will not always be available to serve users. Even apart from this, it is not always practical for a user to contact the producer when he or she needs some information about the meaning or the quality of certain statistical data.

It is important to find ways to motivate producers of statistics to provide good metadata to accompany the statistical data that they produce. Both carrots and sticks are needed. A carrot could be to demonstrate to the producers that there are in fact situations, where even a producer of statistics needs metadata, e.g., when a new statistical survey is going to be designed, and when metadata, e.g., classification categories and their labels, have to be provided to a piece of software. A stick could be a documentation standard that has to be followed. Naturally, such a standard should be supported by a user-friendly tool to make the work as easy as possible for the producer. "Use tools, not rules" is a slogan heard in some statistical offices.
The different metadata holdings and metadata systems that exist in a statistical organization should, ideally, be compatible components of a complete whole, that is, a conceptually and technically well integrated, non-redundant metainformation system that satisfies all important metadata needs of the organization and its users with a minimum of human efforts. In practice, this means that there should be a common conceptual framework and a common technical infrastructure for all metadata holdings and metadata systems. The capturing of certain metadata should

take place when the metadata naturally occur for the first time in a design or production process. The same metadata should not be redundantly captured more than once, and if certain metadata can be derived from already existing metadata, this should be done, and it should be done automatically by means of software tools. Software and applications that need metadata should be able to get them or derive them, as far as possible, from existing sources by means of automatical tools. There should be a kernel of nonredundant metadata from which other metadata can be derived for the different purposes that exist in a statistical organization, and for all important categories of users of statistics, both advanced users like researchers and casual users like journalists and the man in the street.

In other words, in order to facilitate the metadata-related work of statistics producers as much as possible, one should provide tools that facilitate capturing metadata when they first occur and an integrated metadata management system that facilitates the transformation and reuse of existing metadata for other purposes (other stages in the production chain, other software products, other statistical processes.)

Around 1990, Statistics Sweden developed an integrated conceptual framework for systematic and complete descriptions of statistical surveys and statistical observation registers in a broad sense, including registers, statistical production systems based upon administrative sources, and secondary statistical systems like the national accounts. The conceptual framework, called SCBDOK, was developed by Rosén & Sundgren (1991).

The conceptual framework SCBDOK was then used as a basis for the design of a number of metadata holdings and metadata systems at Statistics Sweden.

- A system, also called SCBDOK, was developed for the documentation of final observation registers to be archived for future use by researchers and others. The system is based on a documentation template. Most of the metadata required by the template are free-text metadata, but a subset of the metadata, defined by the METADOK subtemplate, is formalized as relational tables that can also be used automatically by commercial or in-house developed software products for statistics production.
- A standardized quality concept was developed upon the SCBDOK conceptual framework, and this quality concept has been used for producing standardized quality declarations for all official statistics in Sweden. Like the SCBDOK documentations, the quality declaration are structured by means of a templet. As a first step towards more complete, high-quality quality declarations, brief (about 10 pages) product descriptions have been produced, but now the intention is to increase the ambition level.
- With classification theory added to it, SCBDOK has also formed the conceptual basis for the central classification database of Statistics Sweden, intended to cover all national and international standard classifications, including both current and historical versions, as well as Swedish and international versions (of the international classifications).
- SCBDOK, METADOK, the quality declarations, and the classification database are all integrated components of the Internet-based system for dissemination of all Swedish official statistics, ”Sweden´s Statistical Databases”, which became operational 1st of January 1997, and which are now available free of charge (Sundgren 1997).

Furthermore, Statistics Sweden has been the leader of a metadata research project called Integrated MetaInformation Management (IMIM), funded by the European Union within the 4th Framework Programme for Research and Development. Among other things, the IMIM project resulted in a software product called BRIDGE (Rauch & Karge 1999), which is able to accommodate metadata from many different sources and to make metadata available to different software products as well as for different ”human” purposes. The BRIDGE software is based upon an object-oriented data model and an object-oriented database management system called ODABA-2, which turned out to be superior to the state-of-the-art relational data model for the purpose of metadata management. The BRIDGE system is now being used as a base for implementing classification databases in a large number of European countries. A standardized metadata interface called ComeIn has been developed, so as to make it possible to interface metadata holdings based upon software products other than ODABA-2 and BRIDGE.

Statistics Sweden has just undertaken the leadership of another metadata research project, called METAWARE, funded by the European Union within the 5th Framework Programme for Research and Development. This project focuses on metadata management in connection with data warehouses.

More details about the metadata developments at Statistics Sweden can be found in Sundgren (2000).

5. SUMMARY

Metadata is ubiquitous to the processes of producing and interpreting statistics. Defining statistical metadata requires knowledge of the potential users and usages and, thus, it is difficult to do. Its breadth of meaning is such that the metadata producer must address its production in a manner similar to that used for producing the data itself. Moreover, the range of activities included in the cognitive aspects of survey methodology must be extended to metadata production and use. Metadata management must be seen as an integrated part of statistics production, and the metadata management (sub) system itself must be designed from well integrated components, metadata holdings as well as software tools and applications.

6. REFERENCES

Deming, W. E. (1982), *Quality, Productivity, and Competitive Position*, Cambridge, MA: Massachusetts Institute of Technology.

Dippo, Cathryn S. (1997), "Survey Measurement and Process Improvement: Concepts and Integration" in *Survey Measurement and Process Quality*, Lyberg, L., et al. Eds., New York: John Wiley & Sons.

Froeschl, Karl A. (1997), *Metadata Management in Statistical Information Processing*, Vienna: Springer.

Hand, David J. (1993), *"*Data, Metadata and Information*". Statistical Journal of the United Nations Economic Commission for Europe*, 10(2): 143-152. Amsterdam: IOS Press.

Haas, Stephanie (1999), *"Knowledge Representation, Concepts and Terminology: Toward a Metadata Registry for the Bureau of Labor Statistics".* http://ils.unc.edu/~stephani/fin-rept-99.pdf

Hert, Carol (1998), *"Facilitating Statistical Information Seeking On Websites: Intermediaries, Organizational Tools, And Other Approaches."* http://istweb.syr.edu/~hert/BLSphase2.html

Marchionini, Gary and Hert, Carol (1997), *"Seeking Statistical Information in Federal Web Sites: Users, Tasks, Strategies, and Design Recommendations".* http://ils.unc.edu/~march/blsreport/mainbls.html

Marchionini, Gary (1998), "Advanced Interface Designs for the BLS Website: Final Report to the Bureau of Labor Statistics". http://ils.unc.edu/~march/blsreport98/final_report.html

Marchionini, Gary (1999), *"An Alternative Site Map Tool for the Fedstats Statistical Website".* http://ils.unc.edu/~march/bls_final_report99.pdf

Rauch, Lars & Karge, Reinhard (1999*), "BRIDGE - An Object-Oriented Metadata System",* Statistics Sweden & Run Software GmbH, Berlin.

Rosén, Bengt & Sundgren, Bo (1991), *"Documentation for Reuse of Microdata from the Surveys Carried Out by Statistics Sweden"*, Statistics Sweden.

Sundgren, Bo (1973), *"An Infological Approach to Data Bases"*, Doctoral Thesis, University of Stockholm.

Sundgren, Bo (1993), *"Guidelines on the Design and Implementation of Statistical Metainformation Systems".* Revised Version Adopted as *"Guidelines on Modelling Statistical Data and Metadata"* by the United Nations, New York 1995. Also available from Statistics Sweden: R&D Reports 1993:4.

Sundgren, Bo (1997), *"Sweden's Statistical Databases: An Infrastructure for Flexible Dissemination of Statistics".* Report to the UN/ECE Conference of European Statisticians, Geneva. http://www.scb.se

Sundgren, Bo (2000), *"The Swedish Statistical Metadata System"*, Eurostat and Statistics Sweden.

METADATA STANDARDS AND METADATA REGISTRIES: AN OVERVIEW

Bruce E. Bargmeyer, Environmental Protection Agency, and Daniel W. Gillman, Bureau of Labor Statistics
Daniel W. Gillman, Bureau of Labor Statistics, Washington, DC 20212 gillman_d@bls.gov

ABSTRACT

Much work is being accomplished in the national and international standards communities to reach consensus on standardizing metadata and registries for organizing that metadata. This work has had a large impact on efforts to build metadata systems in the statistical community. Descriptions of several metadata standards and their importance to statistical agencies are provided. Applications of the standards at the Census Bureau, Environmental Protection Agency, Bureau of Labor Statistics, Statistics Canada, and many others are provided as well, with an emphasis on the impact a metadata registry can have in a statistical agency.

Standards and registries based on these standards help promote interoperability between organizations, systems, and people. Registries are vehicles for collecting, managing, comparing, reusing, and disseminating the designs, specifications, procedures, and outputs of systems, e.g., statistical surveys. These concepts are explained in the paper.

Key Words: Data Quality, Data Management

1. INTRODUCTION

Metadata is loosely defined as *data about data*. Though this definition is cute and easy to remember, it is not very precise. Its strength is in recognizing that metadata is data. As such, metadata can be stored and managed in a database, often called a *registry* or repository. However, it is impossible to identify metadata just by looking at it. We don't know when data is metadata or just data. Metadata is data that is used to describe other data, so the usage turns it into metadata. This uncovers the weakness of the definition stated above. We need to invoke a *context*, i.e. a point of reference, to identify what we mean by metadata in a given situation. We need to state precisely which data will be used as metadata for our context.

Metadata management refers to the content, structure, and designs necessary to manage the vocabulary and other metadata that describes statistical data, designs, and processes. This includes the development of metadata models to define the content of metadata within some context, building metadata registries to organize the metadata defined in the model, developing statistical terminologies which define and organize terms into a structure with relationships (e.g., a thesaurus), and identifying the relationships between the terminology structure and other metadata and data.

Much work is being accomplished in the national and international standards communities, especially ANSI (American National Standards Institute) and ISO (International Organization for Standardization). to reach consensus on standardizing metadata and registries. This work has had a large impact on efforts to build metadata systems in the statistical community. Several metadata standards are described, and their importance to statistical agencies is discussed. Applications of the standards at the Census Bureau, Environmental Protection Agency, Bureau of Labor Statistics, Statistics Canada, and many others are described. Emphasis is on the impact a metadata registry can have in a statistical agency.

Standards and registries based on these standards help promote interoperability between organizations, systems, and people. Registries are vehicles for collecting, managing, comparing, reusing, and disseminating the designs, specifications, procedures, and outputs of systems, e.g., statistical surveys. These concepts are explained in this paper.

Metadata helps users understand the meaning and quality of data, and registries and the policies put in place for administering them are used to measure and maintain the quality of the metadata. The connection between good metadata and data quality is described, and an overview of procedures for ensuring metadata quality through metadata administration is discussed.

Many people and organizations that plan to implement standards run into a common problem; there are so many standards it is hard to choose the "right" ones. This paper is an attempt to clarify this situation regarding the

management of metadata within the framework of statistical agencies. It is an overview of existing standards that can work together to organize metadata and link them to statistical data and processes.

The paper includes a general description of metadata and metadata registries; a description of metadata management standards; how metadata affects data quality and some measures for quality of metadata itself; and the benefits of implementing metadata standards and registries.

2. STATISTICAL METADATA AND REGISTRIES

2.1 Statistical Metadata

The context we have in mind for metadata in this discussion is statistics. In particular we are interested in the data that are collected and processed through surveys. S*tatistical data*, the data collected and processed through surveys, is called microdata, macrodata, or time series. So, we define *statistical metadata* as the data and documentation that describe statistical data over the lifetime of that data. For the rest of this paper, we will use the term metadata to mean statistical metadata except where noted.

2.2 Metadata Registries

A *metadata registry* is a database used to store, organize, manage, and share metadata. Traditionally, survey groups manage their metadata in their own ways. For example, data dictionaries are created to describe the data elements contained in statistical data sets. There is some coordination of these dictionaries over time but almost no coordination across surveys. A metadata registry is designed to solve this problem by managing metadata from the organization perspective rather than just small program areas. A metadata registry provides for metadata needed to describe objects of interest. It also provides the entities necessary for registration and standardization of those objects (e.g., data elements).

Metadata and metadata registries have two basic purposes (see Sundgren, 1993):

- *End-user oriented purpose*: to support potential users of statistical information (e.g., through Internet data dissemination systems); and
- *Production oriented purpose*: to support the planning, design, operation, processing, and evaluation of statistical surveys (e.g., through automated integrated design and processing systems).

A potential end-user of statistical information needs to identify, locate, retrieve, process, interpret, and analyze statistical data that may be relevant for a task that the user has at hand. The production-oriented user's tasks belong to the planning, design, maintenance, implementation, processing, operation, and evaluation types of activities.

The efficient and effective use of data is made possible by the organized storage and use of metadata. Data sets become much more useful when useful metadata descriptions are readily available. When metadata are centrally maintained for collections of data sets, users who need to determine which files are appropriate for their work can do so. Many types of requests can be answered through metadata queries. Some examples are:

- Which data sets contain specific information, such as yearly income;
- Which data sets share common information from which links can be made to form larger data sets;
- Locate data sets by broad subjects through pointers to specific items under those subjects;
- Monitor data storage system usage by tracking file sizes;
- Locate surveys with similar or a specific set of characteristics.

3. TERMINOLOGY

A *terminology* is a set of *terms*, which in turn is a word or phrase used to designate a *concept*. A *concept* is a unit of thought (see ISO 704). At the most basic level, a terminology is a listing, without structure, or without reference to a specific language. However, concepts must be *defined* in a natural language, possibly employing other terms. The

instantiation of a terminology implies a context. In this case concepts, though not necessarily bound to a particular language, are influenced by social or cultural biases reflected in differences among languages (see ISO 1087-1). Therefore, a terminology may be very specifically related to a subject area, such as retail trade economic surveys.

Terminologies, or classifications, are most useful when a structure is applied to them. The major categories of structure types are thesaurus, taxonomy, and ontology. These are structure types because different kinds of relationships and axioms can be used to build a specific structure for an application. Only the thesaurus structure type has been standardized, but there are draft standards for some types of ontologies. All three structure types are defined below:

- *Thesaurus* - A controlled set of terms covering a specific domain of knowledge formally organized so that the *a priori* relationships between concepts are made explicit;
- *Taxonomy* - A classification according to presumed natural relationships;
- *Ontology* - A formal specification of a conceptualization (i.e., the objects, concepts, and other entities that are assumed to exist in some area of interest and the relationships that hold among them).

Taxonomies are somewhat more restrictive than the other two types. They often contain hierarchical relationships, for example the *North American Industrial Classification System*. However, it is the relationships in a structure that transform a terminology from a simple list to something more meaningful and usable. The breadth and depth of the required relationships help determine the structure type that is needed.

Terminologies help users understand some subject field. The terms and their definitions are the special language that experts in the field use to describe their work and data. Often, though, experts use different terms to mean the same or similar things. In survey statistics, for instance, the terms *frame* and *sample* sometimes mean the same thing. Unless the terms are properly defined, including each of the variants, and useful relationships are established linking the variants and similar terms, confusion will result.

Data elements are the common link across data sets over time and surveys. To understand the content of a data set, one must understand the data elements that make up its data dictionary. Additionally, each survey questionnaire is related to a set of data elements, as are universes, frames, samples, and other design issues.

Terminology is what ties data elements together. Each part of a data element, concept (i.e., definition) and value domain (allowed values), can be described by linking terms to them. The more precisely the terms are defined, the better the terms represent the meaning of the data element. So, users seeking to find data that meets their needs can use terminology to help them find it. A data manager can use terminology to help determine whether data elements mean the same thing, and possibly perform data harmonization. See Figure 1 for the relationship between terminology and other metadata objects.

4. TERMINOLOGY STANDARDS

This section contains descriptions of some standards for terminology. The most general standards apply to thesaurus development, but there are standards for some types of ontologies. The authors are not aware of any standards that describe taxonomies.

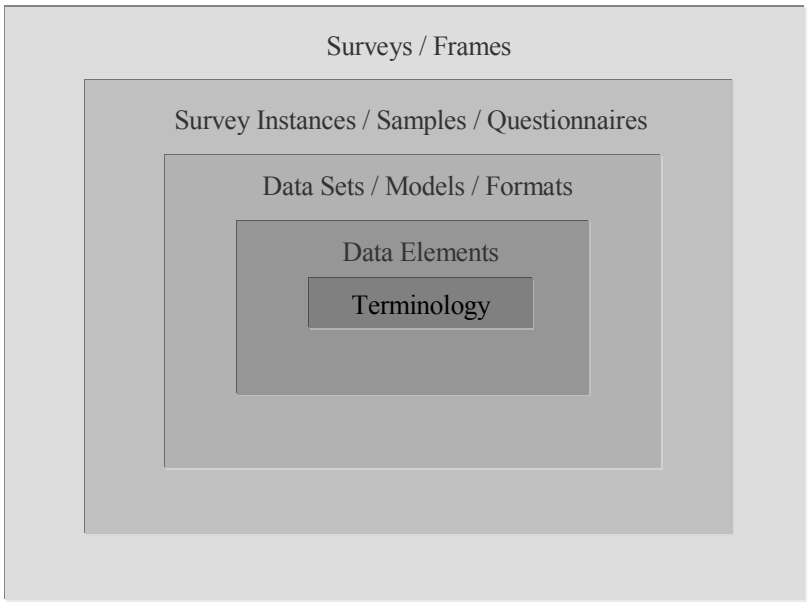
4.1 ISO 704

Principles and methods of terminology. This is a standard about how to write standards for terminology. In a sense, it is the most fundamental of the standards we visit in this section. It is divided into three major sections: *concepts*; *definitions*; and *terms*. These are the basic constructs (described in section 5) that are necessary for terminology. Each is described in more detail in the next paragraphs.

An *object* is an observable phenomenon, and a concept is a mental construct serving to classify those objects. Any object may have multiple concepts associated with it, depending on the context or point of view. A *characteristic* is used to differentiate concepts in a terminology, and different types of characteristics are described. The totality of characteristics for a concept is called its *intension*. The totality of objects sharing all the characteristics of a concept

is its *extension*. *Relationships* are described in some detail because concepts are always related to other concepts in a terminology. Finally, *systems of concepts* are described. A system is the set of concepts of a given subject field. *Definitions* are described in detail. The term is defined, and the purpose of definitions is described. Types of definitions are listed. Within a system of concepts, a definition should be consistent and fix the concept in the

Figure 1: Relationship of Terminology to Metadata



proper position within the system. This is called *concordance*. Detailed principles for developing definitions are described (see also Part 4 of ISO/IEC 11179 below). These principles are general and can be used in many situations. Finally, the use of examples is explained.

Terms are also described in detail. First they are defined. The structure and formation of terms is laid out. This includes constituents (or elements) of terms and methods of forming them. Systems of terms are a coherent set corresponding to the system of concepts the terms represent. The correspondence between a concept and term is described. Emphasis is placed on the difficulty of this in natural language. Requirements for the selection and formation of terms are: linguistically correct, accurate, concise, and some other specialized requirements. Finally, abbreviations are discussed.

4.2 ISO 860

Terminology Work: Harmonization of Concepts and Terms. This standard specifies a methodology for the harmonization of concepts, definitions, terms, concept systems, and term systems. It is a natural extension of ISO 704.

The standard addresses two types of harmonization: *concept harmonization* and *term harmonization*. Concept harmonization means the reduction or elimination of minor differences between two or more closely related concepts. Concept harmonization is not the transfer of a concept system to another language. It involves the comparison and matching of concepts and concept systems in one or more languages or subject fields.
Term harmonization refers to the designation of a single concept (in different languages) by terms that reflect similar characteristics or similar forms. Term harmonization is possible only when the concepts the terms represent are

almost exactly the same. The standard contains a flow chart for the harmonization process and a description of the procedures for performing it.

4.3 ISO 1087-1

Terminology Work: Vocabulary - Part 1: Theory and Application and Terminology Work: Vocabulary - Part 2: Computational Aids in Terminology. These standards establish terminology for doing terminology work. Some of the terms defined in these standards are used throughout this paper. ISO 1087 creates common sets of terms (a special language) for the subject field of terminology application development. Part 1 of the standard identifies, defines, and classifies terms used to develop terminologies for specific applications. The main sections (classes) of the document are: Language and reality; Concepts (includes concepts, characteristics, and relationships); Definitions; Designations (including symbols, names, and terms); Terminology; Aspects of terminology work; Terminological products; and Terminological entries

4.4 Other Standards

There are other standardization efforts for terminology structures and knowledge management. The most important of these are:

- ISO 2788 - Guidelines for the Establishment and Development of Monolingual Thesauri. This standard describes how to build a monolingual thesaurus (i.e., a thesaurus using one language). It uses the theory and structures defined in the standards described above.
- ISO 5964 - Guidelines for the Establishment and Development of Multilingual Thesauri. ISO 5964 sets guidelines for thesauri that have need for more than one language. The content of ISO 2788 is assumed to hold for multilingual thesauri as well as monolingual ones.
- ISO/IEC 14481 - Conceptual Schema Modeling Facility (CSMF).
  This standard defines the constructs to be contained in a modeling facility that can be used to create a formal description of some part of an enterprise. This may include a terminology. A modeling facility is the basis for building a language and a tool to support the activity of creating a formal description. The modeling facility defines the semantics of the language as a set of constructs and how the constructs are related, but not the language syntax.
- NCITS Project 1058 - Knowledge Interchange Format.
- NCITS Project 1059 - Conceptual Graphs

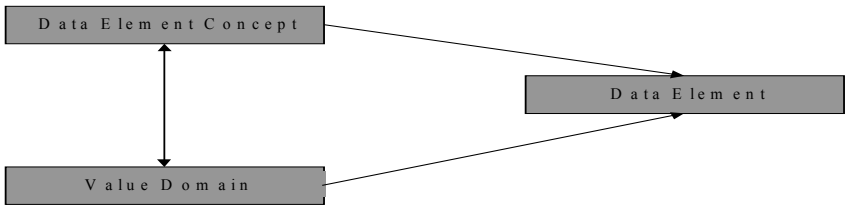


Figure 2: High Level Entity-Relationship Diagram of Metadata Registry

Both of these projects are formal knowledge representation efforts. They and CSMF are part of attempts to standardize ways for a computer to store and use knowledge, described in a formal way (i.e., using first order logic or equivalent schemes). None of these efforts are finished yet, so it is hard to recommend their use. However, they fit into the ontology arena and are part of the larger Artificial Intelligence area of research.

5. ISO/IEC 11179

Specification and Standardization of Data Elements. ISO/IEC 11179 is a description of the metadata and activities needed to manage data elements in a registry. Data elements (or variables) are the fundamental units of data an organization collects, processes, and disseminates. Metadata registries organize information about data elements, provide access to the information, facilitate standardization, identify duplicates, and facilitate data sharing. Data dictionaries are usually associated with single data sets (files or databases), but a metadata registry contains descriptions of the data elements for an entire program or organization.

An important feature of a metadata registry is that data elements are described by a concept (data element concept) and a representation or value domain (set of permissible values). The advantages of this are as follows:

- Sets of similar data elements are linked to a shared concept, reducing search time;
- Every representation associated with a concept (i.e. each data element) can be shown together, increasing flexibility;
- All data elements that are represented by a single (reusable) value domain (e.g. NAICS codes) can be located, assisting administration of a registry;
- Similar data elements are located through similar concepts, again assisting searches and administration of a registry.

Figure 2 is a high level Entity-Relationship diagram (i.e., a model) for a metadata registry of data elements.

Data elements are described by object class, property, and representation. The *object class* is a set of ideas, abstractions, or things in the real world that can be identified with explicit boundaries and meaning and whose properties and behavior follow the same rules. Object classes are the things about which we wish to collect and store data. Examples of object classes are cars, persons, households, employees, orders, etc. However, it is important to distinguish the actual object class from its name. Ideas simply expressed in one natural language (English), may be more difficult in another (Chinese), and vice-versa. For example, "women between the ages of 15 and 45 who have had at least one live birth in the last 12 months" is a valid object class not easily named in English. New object classes are sometimes created by combining two or more other object classes. This example combines the notions of "people between the ages of 15 and 45" with "women who have had live births in the last year".

The *property* is a peculiarity (or characteristic) common to all members of an object class. They are what humans use to distinguish or describe objects. Examples of properties are color, model, sex, age, income, address, price, etc. Again, properties may need to be described using multiple words, depending on the natural language in use.

The *representation* describes how the data are represented (i.e., the combination of a value domain, data type, and, if necessary, a unit of measure or a character set). The most important aspect of the representation part of a data element is the value domain. A *value domain* is a set of permissible (or valid) values for a data element. For example, the data element representing annual household income may have the set of non-negative integers (with units of dollars) as a set of valid values. This is an example of a *non-enumerated domain*. Alternatively, the valid values may be a pre-specified list of categories with some identifier for each category, such as:

1	$0 - $15,000
2	$15,001 - $30,000
3	$30,001 - $60,000
4	$60,001 - +

This value domain is an example of an *enumerated domain*. In both cases, the same object class and property combination - the annual income for a household - is being measured.

The combination of an object class and a property is a *data element concept* (DEC). A DEC is a concept that can be represented in the form of a data element, described independently of any particular representation. In the examples

above, annual household income actually names a DEC, which has two possible representations associated with it. Therefore, a data element can also be seen to be composed of two parts: a data element concept and a representation.

Figure 3, below, illustrates the ideas discussed above. It shows that each data element concept is linked to one or more data elements; an object class may be generated from other object classes; a data element concept has one object class and one property associated with it; and a data element has one data element concept and one representation associated with it.

ISO/IEC 11179 - Specification and Standardization of Data Elements - is divided into six parts. The names of the parts, a short description of each, and the status follow below:

- Part 1 - *Framework for the Specification and Standardization of Data Elements* - Provides an overview data elements and the concepts used in the rest of the standard. This document is an *International Standard*.
- Part 2 - *Classification of Data Elements* - Describes how to classify data elements. This document is an *International Standard.*
- Part 3 - *Basic Attributes of Data Elements* - Defines the basic set of metadata for describing a data element. This document is an *International Standard*. It is currently being revised.
- Part 4 - *Rules and Guidelines for the Formulation of Data Definitions* - Specifies rules and guidelines for building definitions of data elements. This document is an *International Standard*.
- Part 5 - *Naming and Identification Principles for Data Elements* - Specifies rules and guidelines for naming and designing non-intelligent identifiers for data elements. This document is an *International Standard*.
- Part 6 - *Registration of Data Elements* - Describes the functions and rules that govern a data element registration authority. This document is an *International Standard*.

The revision of ISO/IEC 11179-3 (Part 3) will include a conceptual model for a metadata registry (for data elements). The metamodel, or metadata model for data elements, provides a detailed description of the types of information which belong to a metadata registry. It provides a framework for how data elements are formed and the relationships among the parts. Implementing this scheme provides users the information they need to understand the data elements of an organization. Figure 2 is high level view of the metamodel.

There are two additional important features of the metamodel. It provides for the *registration* of metadata. Registration is a process that administers metadata. It keeps track of who submitted the metadata, who is responsible for it, and the quality of the metadata provided. The ability to measure metadata quality is very important and is an issue which is often overlooked. The other important feature is that the metamodel contains a common metadata entity called *administered_ component*. This entity captures the metadata common to all objects described in the registry. From the perspective of the common metadata, the *administere_ component* entity is like a library card catalog. This makes searching for some objects much easier.

Finally, the revision of ISO/IEC 11179-3 (Part 3) will be more than a data element standard; it will be a metadata registry standard. Many agencies, including the U.S. Census Bureau, are implementing these standards as part of their metadata repository design efforts.

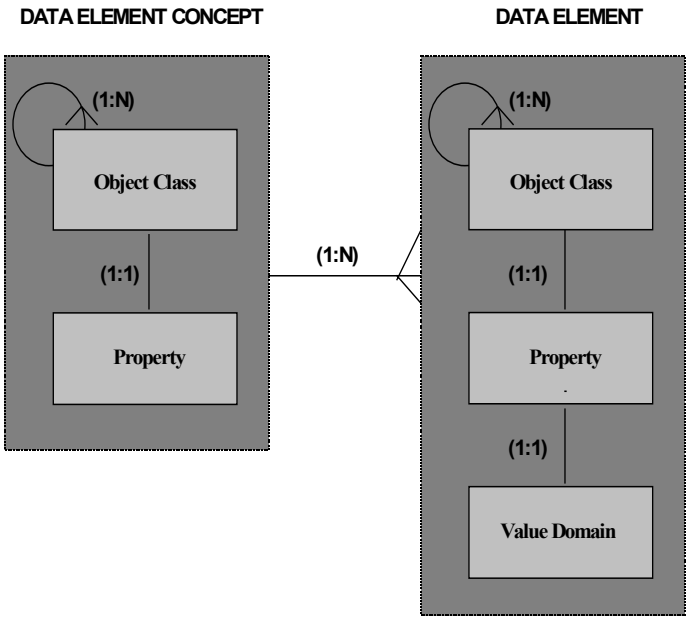


Figure 3: Fundamental Concepts of Data Elements

6. IMPLEMENTATIONS OF ISO/IEC 11179

Many organizations are implementing metadata registries based on ISO/IEC 11179. This section contains descriptions of several of these efforts.

6.1 The Intelligent Transportation System Data Registry Initiative

The Department of Transportation is working on developing 50 to 60 standards for interoperability among the systems that will comprise the nation's Intelligent Transportation System. Five Standards Development Organizations are cooperating in the development of these standards, in addition to identifying a controlled vocabulary and various access paths for information. A major project within this initiative is the Intelligent Transportation System (ITS) Data Registry. It is being designed and developed by IEEE[1] and is based upon the ISO/IEC 11179 standard's concepts. The Transportation Department's explicit use of ISO/IEC 11179 as part of their standards program for the intelligent highway projects has been for the purpose of encouraging and enabling the transportation agencies of the 50 States and Territories to be able to exchange and work with data that has consistent semantics across the various governmental and other organizations involved.

6.2 The Environmental Data Registry

The Environmental Protection Agency (EPA) is developing methods to: a) share environmental data across program systems; b) reduce the burden of reporting regulatory and compliance information; c) improve access to environmental data via the web; and d) integrate environmental data from a wide variety of sources. To accomplish these improvements, the Environmental Information Office is creating the Environmental Data Registry (EDR). The EDR is the Agency's central source of metadata describing environmental data. In support of environmental data standards, the EDR offers well-formed data elements along with their value domains. The EDR is also a vehicle for reusing data elements from other data standards-setting organizations.

The EPA Administrator established the EDR as the agency resource for describing and standardizing environmental data. The EDR supports several strategic goals of EPA, including One Stop reporting, the Reinvention of Environmental Information, and the Public Right to Know. It is used for describing environmental data found inside

[1] Institute of Electronics and Electrical Engineers

and outside of EPA. It is also used by state environmental cleanup program offices to facilitate sharing data among themselves - data that are held only by states and not reported to EPA. The EDR is used to record the results of discussions that rage between program offices about data content and design. It is populated with metadata describing a wide spectrum of environmental data including data in environmental information systems, environmental Electronic Data Interchange (EDI) messages, an environmental data warehouse, environmental regulations, etc. Well-formed data are registered for voluntary use. Mandatory data standards are registered for Agency-wide implementation. The EDR is accessible from the World Wide Web and each month serves up hundreds of thousands of pages. Users download metadata for data elements and groups of data elements. Users also download the entire registry contents.

The registry initiative is engaging European environmental organizations for joint U.S. and European sharing of worldwide environmental data. A major EPA effort is now underway to work on terminology. EPA is building a prototype terminology reference system, as a module of the EDR, that will be compatible with the European environmental terminology system, the General Multilingual Environmental Thesaurus (GEMET). The EDR system provides a direct link to the prototype EPA Terminology Reference System (TRS). The TRS is a new web-based tool for managing lists of terms to be used in data elements, classification systems, ontologies, as well as in text web pages, and other electronic records. Currently, the TRS houses the General European Multilingual Environmental Thesaurus (GEMET).
See: http::/www.epa.gov/edr, http://www.epa.gov/trs and http://www.epa.gov/crs

6.3 Australian National Health Information Knowledgebase

The Australian Knowledgebase is an electronic storage site for Australian health metadata, and includes a powerful query tool. You can use the Knowledgebase to find out what data collections are available on a particular health-related topic or term, and any related official national agreements, definitions, standards and work programs, as well as any linked organizations, institutions, groups, committees or other entities. The Knowledgebase provides direct integrated access to the major elements of health information design in Australia:

- The National Health Information Model;
- The National Health Data Dictionary;
- The National Health Information Work Program; and
- The National Health Information Agreement.

The Knowledgebase does not, as a rule, provide access to actual data through its searching and querying tools, but it is a planned future development.
See: http://www.aihw.gov.au/services/health/nhik.html

6.4 The United States Health Information Knowledgebase

The Department of Defense - Health Affairs (HA) in collaboration with the Health Care Financing Administration (HCFA) is developing the United States Health Information Knowledgebase (USHIK) Data Registry Project. The project goal is to build, populate, demonstrate, and make available for general use a data registry to assist in cataloging and harmonizing data elements across multi-organizations. The requirements team includes representatives from the Department of Veteran Affairs, the Health Level Seven (HL7) standards committee, the Health Care Financing Administration, and the Department of Defense Health Affairs office. The implementation builds on Environmental Protection Agency and Australian Institute of Health and Welfare implementations and utilizes DoD - Health Affairs' Health Information Resource Service (HIRS) to develop and implement a data registry. The project utilizes selected Health Insurance Portability and Accountability Act (HIPAA) data elements for demonstration. The data elements are those used in standards by the X12 (EDI[2]) standards committee, the HL7 standards committee, the National Council of Prescription Drug Program (NCPDP), and the National Committee on Vital and Health Statistics (NCVHS).

[2] Electronic Data Interchange

An EPA, HCFA, and DoD-HA joint effort also is an initial model for interagency agreements and working arrangements can be made between agencies for synergistic development of metadata and data registry technology.
See: http://hmrha.hirs.osd.mil/registry/

6.5 The Census Bureau Corporate Metadata Repository

The Census Bureau is building a unified framework for statistical metadata. The focus of the work is to integrate ISO 11179 and survey metadata, using the metadata to enhance business applications. A production corporate metadata registry (CMR) is under development based on an extended model including ISO 11179 and a business data model. The goal is put metadata to work to guide survey design, processing, analysis, and dissemination.

The development process features pilot projects that use the CMR from different subject areas within the agency. Each of these projects focuses on a different aspect of the CMR and the information potential it provides. As more projects use the CMR, its visibility and usefulness within the agency will increase.

Current project applications include the American Fact Finder - Data Access and Dissemination System. This project is a large effort to disseminate Decennial Census, Economic Censuses, and American Community Survey data via the Internet. Other projects are integrating the CMR with electronic questionnaire design tools, batch and interactive CMR update and administration tools, and interfaces with statistical packages such as SAS.

6.6 Statistics Canada Integrated MetaDataBase

Statistics Canada is building a metadata registry, called the Integrated MetaDataBase, based on the same conceptual model for statistical metadata developed at the Census Bureau. This effort is still in the design and initial implementation stages. It will integrate all the surveys the agency conducts, contain many standardized and harmonized data elements, and link statistical data to the survey process.

6.7 OASIS and XML.org XML Registry and Repository Work

OASIS, the Organization for the Advancement of Structured Information Standards is developing a specification for distributed and interoperable registries and repositories for SGML[3] and XML[4] DTDs (Document Type Definitions) and schemas. XML.org, an initiative within OASIS, is implementing one such registry and repository. The basis for much of the specification is ISO/IEC 11179.
See: http://www.oasis-open.org/html/rrpublic.htm

7. REGISTRATION AND QUALITY

The quality of data is enhanced when the proper metadata is available for that data. Data is more understandable and useable in this case. Also, data quality statements themselves are metadata. So, metadata describing sampling errors, non-sampling errors, estimations, questionnaire design and use, and other quality measures all need to be included in a well-designed statistical metadata registry. However, this does not say anything about the quality of the metadata itself.

Registration is the process of managing metadata content and *quality*. It includes:

- making sure mandatory attributes are filled out;
- determining that rules for naming conventions, forming definitions, classification, etc. are followed;
- maintaining and managing levels of quality.

Registration levels (or statuses) are a way for users to see at a glance what quality of metadata was provided for objects of interest. The lowest quality is much like "getting some metadata"; a middle level is "getting it all" (i.e., all that is necessary); and the highest level is "getting the metadata right".

[3] Standard Generalized Mark-up Language
[4] eXtensible Mark-up Language

Semantic content addresses the meaning of an item described by metadata. Usually the name and definition are the extent of this, but for statistical surveys, much more relevant information is necessary to describe an object. A data element has a definition, but additional information that is necessary to really understand it is: the value domain; the question that is the source of the data; the universe for the question; the skip pattern in the questionnaire that brought the interviewer to the question; interviewer instructions about how to ask the question; sample design for the survey; standard error estimates for the data; etc. The model driven approach to the CMR is a start to understanding this.

Once the semantic content is really known, then the work to harmonize some data across surveys and agencies can begin. Harmonization can occur at many levels, e.g., data, data set, and survey.

Metadata quality is a subject that has received much less attention than content and organization. Metadata has quality when it serves its purpose - allows the user to find or understand the data which is described. As such, metadata quality has several dimensions:

- the full set of metadata attributes are as complete as possible;
- the mandatory metadata attributes describe each object uniquely;
- naming conventions are fully specified and can be checked;
- guidelines for forming definitions are fully specified and can be checked;
- rules for classifying objects with classification schemes are specified and can be checked;
- classification schemes are as complete as possible.

Research will focus on ways of measuring quality using these and other criteria.

8. BENEFITS

The benefits of implementing an ISO/IEC 11179 metadata registry are several:

- Central management of metadata describing data and other objects throughout the agency;
- Increased chances of sharing data and metadata with other agencies that are also compliant with the standard;
- Improved understandability of data and survey processes for users;
- Single point of reference for data harmonization;
- Central reference for survey re-engineering and re-design.

The structure of an ISO/IEC 11179 metadata registry has many points at which terminology will aid in searching and understanding the objects described. As described above, terminology is developed to aid the user, designer, and analyst. The *meaning* of an object is clarified through the set of all the terms linked to it. Although a definition is important for understanding an object, it often does not convey the full context in which the definition is made. A good example is a question in a questionnaire. The question wording itself serves as its definition, but the universe for which the question is asked or about is usually not specified. That context is inferred from the flow and answers to previous questions. However, appropriate terms associated with a question can convey some of this necessary information without resorting to following a complicated questionnaire down to the question under consideration.

In conclusion, many organizations are implementing ISO/IEC 11179 metadata registries, including the U.S. Census Bureau, U.S. Environmental Protection Agency, U.S. Health Care Financing Administration, Australian Institute for Health and Welfare, and others. The international standards committee **ISO/IEC JTC1/SC32/WG2** (Metadata) is responsible for developing and maintaining this and related standards. Participation by national statistical offices through the appropriate national standards organization will make this effort much stronger and provide a means to interoperability across national boundaries for statistics.

9. REFERENCES

- Gillman, D. W., Appel, M. V., and Highsmith, S. N. Jr. (1998), “Building a Statistical Metadata Repository”, Presented at METIS Workshop, Geneva, Switzerland, February, 1998.
- Graves, R. B. and Gillman, D. W. (1996), "Standards for Management of Statistical Metadata: A Framework for Collaboration", ISIS-96, Bratislava, Slovakia, May 21-24, 1996.
- ISO 704, Principles and Methods of Terminology, 1987, International standard.
- ISO 860, Terminology Work - Harmonization of Concepts and Terms, 1996, International standard.
- ISO 1087-1, Terminology Work - Vocabulary - Theory and Application, 1995, International standard.
- Sundgren, B. (1993), "Guidelines on the Design and Implementation of Statistical Metainformation Systems", R&D Report Statistics Sweden, 1993:4.
- Sundgren, B., Gillman, D. W., Appel, M. V., and LaPlant, W. P. (1996), "Towards a Unified Data and Metadata System at the Census Bureau", Census Annual Research Conference, Arlington, VA., March 18-21, 1996.

USE OF METADATA FOR THE EFFECTIVE INTEGRATION OF DATA FROM MULTIPLE SOURCES

Mark E. Wallace and Samuel N. Highsmith, United States Census Bureau*
Mark E. Wallace, United States Census Bureau, EPCD, Room 2570-3, Washington, DC 20233
Mark.E.Wallace@ccmail.census.gov

ABSTRACT

As survey practitioners in statistical agencies continue to expand the collection and dissemination of data electronically, the statistical community has embraced the notion that using data integrated from multiple sources is more powerful than relying on stand-alone data sets. What has been termed integrated statistical solutions (ISS) consists of providing data users with answers to their questions, without the user first having to know the structure of the government or agency or how the data files are organized or formatted. Given the trends in technology and user expectations, the movement toward producing such integrated data applications is certainly inevitable. As a result, the role of metadata to support applications using integrated statistical information has become increasingly important. This paper compares and contrasts alternative metadata-sharing support structures that statistical agencies might employ to enable the integration of data and metadata from multiple sources.

Key Words: Repositories, Registries, Metadata-sharing, FedStats, ISS

1. INTRODUCTION

As survey practitioners in statistical agencies continue to expand the collection and dissemination of data electronically, the role of metadata as a driver and in support of understanding the content of our statistical information is becoming increasingly important. We have entered this new century amidst a wave of technological innovation, featuring the Internet and the World Wide Web as the primary means of information dissemination. At the same time, the international statistical community is being challenged with an urgent and critical need for more accurate, timely, relevant, accessible, and interpretable data. Several thought leaders in various statistical agencies around the world are attempting to meet these user needs by working together, like never before, to implement a modernized, customer-driven, cross-program, and cross-agency integrated data access and dissemination service capability.

Over the past few years, the statistical community has embraced the notion that using data integrated from multiple sources is more powerful than relying on stand-alone data sets. What has been termed integrated statistical solutions (ISS) consists of providing data users with answers to their questions, without the user first having to know the structure of the government or agency or how data files are organized or formatted. And, given the trends in technology and user expectations, the movement toward producing integrated data applications is certainly inevitable. However, the question now before us is "***How*** will we realize the vision?" The answer lies in the development of the statistical and spatial metadata infrastructure support to drive the integration of data and metadata from multiple agencies.

2. BACKGROUND

Currently there is no assurance that similarly named items will be referring to the same thing. Conversely, there are likely to be data elements with different names in different databases that are actually descriptive of the same things. Further, there is currently a decided lack of rules for identifying which statistics are "integrateable" for various levels of geography, time, and topic. Hence, it is evident that for effective comparison of the meanings and intelligent use of information items in various statistical databases, metadata naming and attributes systems need to be based on international standards, such as the multi-part ISO/IEC 11179, "Specification and Standards of Data Elements".

A number of federal agencies have developed or are developing their own metadata registries and are using these systems for operational needs. These include the Environmental Protection Agency, the Bureau of the Census, the Department of Transportation, the Department of Energy, the Bureau of Labor Statistics, the Health Care Financing Administration, the Department of Defense – Health Affairs, and the Department of Veterans Affairs. Operationalizing metadata

---

* This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review by the Census Bureau than its official publications. This report is released to inform interested parties and to encourage discussion.

repositories in these agencies is certainly a good first step. Next, however, to support ISS, agencies must collaborate to develop methods that support the automated sharing of metadata between agencies that have metadata registries.

This paper describes some important initial steps that agencies and researchers might take to initiate the development of these metadata-sharing methods. These steps include efforts by a team of several federal agencies, working with partners in the data user and research communities, to develop models that demonstrate the value of, as well as the technical capabilities for, dynamically integrating data from multiple sources.

Specifically, over the next several months, this work will involve experimenting with various combinations of hardware, software, data and metadata from the various partners collaborating on this project. There is the potential for determining and validating best approaches for sharing data and metadata from multiple sources across agencies. We hope that lessons we learn from this work will help foster the ushering in of a data and metadata-sharing environment from which a number of new cross-agency topical and geographic-related data access and dissemination tools and products can be introduced.

3. FIRST STEPS

Documenting data sets is a first step in enabling data sharing between agencies. Statistical and geographic databases have been built thus far to support the mandates of single institutions or parts of an institution. All who collect and manage data for activities related to their own responsibilities will need to understand and appreciate the value of those data to others and to collect and structure their data accordingly. To take full advantage of the opportunities offered by new technologies, business, government, and academia will need to develop, support, and fund metadata on a systematic and ongoing basis as well as promote access for all.

The Internet has clearly changed expectations and heightened knowledge about the ease of access to information as well as broadened the universe of users. Customers, both internal and external, associated with the various government agencies expect technology to provide easy and rapid access to documented, usable, and interpretable data. Developing metadata repositories based on international standards will enable us to address the increasing demand for rapid access to documented and usable data. This might eventually be expandable to a virtual government-wide, or even global, level. Metadata would be the integrating mechanism for providing government-wide information.

Presently, at the Census Bureau, business processes touch metadata from the survey/census design through the dissemination life cycle (Gillman and Appel 1999). Yet, metadata is not formally cataloged until the dissemination phase where it is required for Bureau dissemination systems and products. We are now beginning to see the advantage in cataloging these data at each phase of the survey life cycle so that metadata doesn't have to be recreated during the dissemination phase. This approach also provides value added during each of the core business processes.

4. INTERAGENCY COLLABORATIVE PROJECTS

Various teams composed of staff from all directorates of the Census Bureau in collaboration with the State Data Centers, FedStats agencies, and the research community are researching the potential and validating data integration processes for accessing and integrating both micro and macro level datasets (and their metadata) that are Census Bureau-based and/or remote. These teams also are developing data integration tools to create profiles (predefined and user-specified), and to give customers new data integration functionality, providing information based solutions not currently available in our existing data access tools. It will be through metadata that much of this work will be accomplished.

These projects will make use of a collaborative, multi-agency laboratory server environment that is being established to evaluate different tools and approaches, use corporate metadata policies and techniques, and include profiles with some level of graphic display capabilities.

5. COMPARISON OF VARIOUS METADATA MODEL APPROACHES FOR INTEGRATING DATA FROM MULTIPLE SOURCES

In the course of developing our own Corporate Metadata Repository (CMR) at Census, and in working with other organizations to experiment with ways to share and integrate disparate data sets, we have examined the feasibility of

implementing three possible metadata models. Our goal is to determine which one(s) might best, over time, bring about the desired outcome of effectively and intelligently integrating data from multiple sources. To this end, we have compared and contrasted a number of factors with regard to each approach.

The initial evaluation was performed by a cadre of Census staff from various areas. This group developed the factors used to evaluate the three models and ranked each model numerically. Once their initial evaluation was complete, the metadata model approaches and their preliminary evaluations were shared with representatives of other agencies which are doing work in this area. Their feedback is incorporated in the evaluations below.

The factors we examined are *performance, scalability, cost, interoperability, control, flexibility, security, short and long-term implementability, maintenance, and reliability.* For each metadata model, we assigned two values (on a scale of 1 to 10 with 1 being the worst and 10 being the best) to each factor. The best possible score for the 10 factors examined would be 100. These two values are based on the experiences of a number of agencies *now* and what we project will be happening *in 5 years*.

5.1. Single Source Repository for Metadata (SSRM)

This model stores all the agreed upon metadata elements for all participating agencies in one central repository. All applications would use it directly for access and maintenance. This would require all participating agencies to agree on a core set of metadata elements. It is the easiest model to build from the standpoints of interoperability, short-term implementability, control, security, maintenance, and reliability. The difficulties in this model occur in the areas of performance, scalability, cost, flexibility, and long-term implementability. In fact, this approach is the very methodology used in many of our existing stovepipe systems. This would require construction and continuing support of one central repository to be used by all participating agencies. This approach would not support the concept of unique or different metadata element requirements. All agencies would have to fit the metadata for their various data sets into one standard definition. And, as more agencies participate over time, more resources to support this SSRM would be required by the hosting agency.

- *Performance – As usage by multiple organizations grows, reasonable performance of one large application becomes increasingly difficult to provide.*
- *Scalability – This system has limited scalability due to the increasing cost of providing reasonable performance with a central resource.*
- *Cost – The initial cost is very reasonable. However, this model becomes very expensive as additional usage and more requirements are placed on it.*
- *Interoperability – This is a very interoperable system with a single applications program interface for application interface.*
- *Control – This implementation is the easiest to control.*
- *Flexibility – Mediocre flexibility due to difficulty in changing metadata elements. Since every organization using this system must provide a strictly agreed upon set of metadata elements, adding new elements or changing existing elements is very difficult.*
- *Security – This implementation, by being centralized, is the most secure.*
- *Short and Long-term Implementability – For the short term, this is the easiest model to implement. Over the long term, this approach is less viable since implementability is likely to decrease in proportion to the number of participating agencies.*
- *Maintenance – Very maintainable.*
- *Reliability – With fewer points of failure, this model should provide the highest reliability.*

5.2. Distributed Dissemination Metadata Repository (DDMR)

This model entails each agency providing an agreed upon core set of metadata elements, but they would not necessarily conform to one international metadata standard. Instead, metadata sharing among the agencies would be accomplished via the use of a standardized software supported metadata interchange. In this model, each organization would build its own metadata repository using the agreed upon metadata elements and underlying model. The metadata elements would be limited to those required by data dissemination applications, meaning that documentation of the survey process not

needed for dissemination would not be provided. Metadata added to one repository would normally be replicated to all other repositories using a "token" registration scheme to ensure all registration efforts succeed. Organizations would be able to be selective on which other organizations their metadata would be sent to. They could also choose not to share some or all metadata. Each organization would be responsible for providing their own metadata registry integrity, to include registry backup operations. The end result would be the creation of many unique metadata registries using a common registry format and shared software tools. Since we would need a complete set of shared functionality for this approach, it assumes a common architecture used by all. To implement this model, some organization would have to develop, distribute, and support the application software.

- *Performance – Since each organization will support and access its own metadata registry, performance is the responsibility of each organization. Access to distributed data residing in other organizations is directly related to the speed of the internet connection deployed by each organization.*
- *Scalability – This system has somewhat limited scalability due to the technique of distributed data and multiple unique metadata registries. As various agencies share increasing amounts of metadata, the size of the DDMR will increase exponentially.*
- *Cost – The cost is reasonable, especially since each organization shoulders the cost of its own metadata registry. The only really expensive part of this implementation is the construction of metadata and support of a shared set of software to allow distributing metadata across registries.*
- *Interoperability – This is a very interoperable system when the shared software is developed and distributed. The unfortunate side effect of this approach is that it could become obsolete when metadata interchange standards come into existence unless the software development effort is tied to the standards development effort.*
- *Control – Each organization can establish and control its own security, which means control is only as good as each organization's implementation.*
- *Flexibility – Good flexibility in changing metadata elements.*
- *Security – This implementation is only as secure as the participating organizations make it.*
- *Short and Long-term Implementability – In the short term, this system requires fairly sophisticated software development and support by a lead organization willing to take on responsibility for the software, and implementation and support of the model. For the long term, this system will become increasingly difficult to support as the number of participating organizations increases. In addition, the lead organization will be burdened with the responsibility of updating and distributing application software and tools over time.*
- *Maintenance – Very much dependent on each organization deploying it.*
- *Reliability – Very much dependent on each organization deploying it.*

5.3. Federation of Unique but Related Metadata Repositories (FMR)

This model is based on the concept of a logically central metadata repository (Sundgren, et al 1996). Each agency would maintain their own registry, including the agreed upon set of metadata elements, and, when it becomes available, each agency would conform to one international standard. Recognizing that needs vary across agencies, agencies could, in effect, build their own metadata mart as long as their registries were in compliance with the standard. This way, agencies would be able to extend metadata requirements beyond the core set, add application specific metadata or tune their implementation to meet their application performance requirements.

This is the basis for the model currently under construction at the U.S. Census Bureau (Gillman and Appel 1999). The core component is a centrally built and maintained metadata repository like the SSRM. This metadata repository and its supporting tools can be used directly by participating organizations, should they so choose, using the agreed upon metadata elements. However, to support unique organizational requirements, a second technique is for departments to build and support their own metadata repository. In this more loosely coupled setup, the organization can copy the central metadata repository and make any additions required by their applications.

Another technique, particularly useful where an organization has already put in place their own metadata repository, is to map the agreed upon metadata elements of the central repository to the components of the FMR and build an interchange. This particular flavor of the FMR approach envisions development and use of a standard metadata interchange format to exchange metadata between repositories. Although this international metadata standard does not yet exist, it appears likely that such a standard based on XML is likely to emerge in the next few years.

- *Performance – Performance of the actual central system is relatively easy to provide. Performance of distributed repositories is very much under the control of the separate organization building and supporting their own registry.*
- *Scalability – This system has virtually unlimited scalability due to the technique of distributed data and multiple unique metadata registries.*
- *Cost – The cost is reasonable, especially since each organization shoulders the cost of its own metadata registry. The only really expensive part of this implementation is the construction and support of a shared set of software to allow distributing metadata across metadata registries.*
- *Interoperability – This is a very interoperable system when the shared software is developed and distributed. The unfortunate side effect of this approach is that it could become obsolete when metadata interchange standards come into existence unless the software development effort is tied to the standards development effort.*
- *Control – Each organization can establish and control its own security, which means control is only as good as each organization's implementation.*
- *Flexibility – This is the most flexible model to implement.*
- *Security – This implementation is only as secure as the participating organizations make it.*
- *Short and Long-term Implementability – For the short term, this is a very easy model to implement. For the long term, this approach continues to be viable.*
- *Maintenance – Very much dependent on each organization deploying it.*
- *Reliability – Very much dependent on each organization deploying it.*

6. SUMMARY OF SCORES FOR EACH METADATA MODEL

Based on the above evaluations, none of the three models demonstrated overwhelming superiority over the others. Over the long term however, the FMR – which is a hybrid approach – appears to be the best. Below are summary evaluations of each model along with numerical scores.

6.1. Single Source Repository for Metadata (SSRM)

The SSRM represents the most centralized approach to metadata sharing. Because all of the dissemination metadata would be in one repository, interoperability and control are optimized. Maintenance and reliability are also highly rated since all of the shared metadata would be maintained together, and would not be subject to the unique requirements of the individual participating data and metadata suppliers. With a small group of participating organizations, this model should be very implementable. As the system grows over time however, and the number of users increases, maintaining a reasonable level of performance will become expensive and difficult to manage.

Also, a centralized approach requires data and metadata suppliers to adjust their metadata standards – at least for the core elements – to fit within the requirements of the single source repository. This limits flexibility and makes implementability less feasible over the long term unless an international standard were adopted and adhered to.

6.2. Distributed Dissemination Metadata Repository (DDMR)

The DDMR model provides the least centralized approach to metadata sharing. Since each participating organization would be responsible for establishing and maintaining its own metadata registry, this model is considered to be infinitely scaleable. Costs associated with developing a virtual web of metadata registries would also be borne by individual organizations which is considered to be a strength. However, control, security, maintenance, and reliability of component registries would remain under the purview of the organization sponsoring the registry, which could become an excessive burden.

In addition, the sponsoring organization will need to take responsibility for developing and maintaining the underlying architecture for the distributed environment and fairly sophisticated data access and integration tools which would be common to all registries. Over the long term, this would also become burdensome, and in fact may well become impossible as more and more agencies use increasingly larger replicated repositories.

6.3. Federation of Unique but Related Metadata Repositories (FMR)

The FMR represents a hybrid approach which proposes to take the best functionality from both the centralized and the distributed models, and implement a very flexible model based on international standards for metadata registries and interchange format. As a result, ratings for most evaluation factors are equal to or better than the scores of the other models. Notable exceptions are, interoperability, control, maintenance and reliability, where the centralized model is strongest. This is due to the fact that metadata registries are developed and maintained individually by each sponsoring organization notwithstanding the logically central repository to which they all contribute. The expectation that there will be variability in these areas is considered a weakness that may be expected to diminish somewhat over time. For this reason, the FMR, while scoring highest of the three approaches over the long term, is not significantly superior especially in the short term. Nevertheless, the FMR model is probably the best choice because of its long term superiority.

6.4. Numerical Scores

As part of our analysis, the three models were subjected to numerical ratings for each of the factors. The highest achievable score was 100. Below are the scoring results by evaluation factor – both short term and long term.

Evaluation Factors	**Score**					
	SSRM		**DDMR**		**FMR**	
	Now	*In 5 yrs.*	*Now*	*In 5 yrs.*	*Now*	*In 5 yrs.*
Performance	4	2	5	5	4	6
Scalability	4	2	6	3	10	10
Cost	7	3	7	7	7	7
Interoperability	10	10	8	4	8	6
Control	10	10	5	5	6	7
Security	7	7	5	5	6	7
Implementability	10	7	6	2	9	9
Flexibility	4	3	7	7	10	10
Maintenance	9	9	5	5	5	7
Reliability	9	9	5	5	5	7
Total Score	74	63	59	48	70	76

7. NEXT STEPS

Based on an examination of important factors with regard to the various metadata models for sharing data and metadata from multiple sources, it appears that a most viable approach may be to adopt the Federation of Unique but Related Metadata Repositories (FMR) model. We think that this system offers the most flexibility not only now but well into the future. When an international metadata interchange format becomes available, this system will be well positioned to use and take advantage of it.

Some of the questions we hope to be able to answer over the next few years are:

1. International standards for metadata registries will exist soon. What will be the impact of new technologies including XML?
2. How well will we be able to adopt standards within individual organizations, and develop/maintain the necessary vision to support data access and integration capability within and across agencies?
3. Will we receive the support necessary to continue collaborative efforts among federal agencies, academia and the research community?
4. How accurate are our predictions regarding the choice of an appropriate metadata sharing model to develop? We are currently researching the feasibility of developing the DDMR and FMR models. If research show that we are incorrect in our assumptions concerning the scalability and implementability of the DDMR, it could prove to be a viable model.

We will certainly learn more about metadata sharing efforts such as the FedStats Product Concepts Working Group as we continue to collaborate with various parties in the public and private sectors and in the academic and research communities.

REFERENCES

Capps, Cavan P., Green, Ann, and Wallace, Mark E. (1999), "The Vision of Integrated Access to Statistics: the Data Web", paper submitted to the Association of Public Data Users.

Gillman, Daniel W. and Appel, Martin V. (1999), "Statistical Metadata Research at the Census Bureau", *Proceedings of the Federal Committee on Statistical Methodology Research Conference*, pp.1-10.

ISO/IEC 11179 (1994-2000), *Information Technology – Specification and Standardization of Data Elements*, Parts 1-6, Draft International Standards.

Schneider, Daniel (1999), "Information and Database System Interoperability with Assured Semantic Consistency: The Role of Data Semantics Management Systems – A White Paper for Federal Agency CIOs and IT Architects" (Draft).

Sundgren, B., Gillman, D.W., Appel, M. V., and LaPlant, W. P. (1996), "Towards a Unified Data and Metadata System at the Census Bureau", *Proceedings of the Census Annual Research Conference*.

Wallace, Mark E., Landman, Cheryl M., Sperling, Jon, and Buczinski, Carla (1999), "Integrated Information Solutions - The Future of Census Bureau Data Access and Dissemination", *Proceedings of the Government Statistics, Social Statistics, Survey Research Methods Section, American Statistical Association*.

Wallace, Mark E. (2000), "User Driven Integrated Information Solutions - Digital Government by the People for the People", *Topic iv: Improving data dissemination strategies, Seminar on Integrated Statistical Information Systems (ISIS 2000).*

STATISTICAL METADATA: THE REAL WORLD OF IMPLEMENTATION

Michael Colledge, OECD and Ernie Boyko, Statistics Canada
Michael Colledge, Statistics Directorate, OECD, 2 rue Andre Pascal, 75775 Paris Cedex 16, France
michael.colledge@oecd.org
Ernie Boyko, Statistics Canada, 2-O, RH Coats Building, Ottawa, Ontario, Canada, K1A OT6
eboyko@statcan.ca

ABSTRACT

The management of metadata and classification of objects described by metadata are major problems confronting the successful implementation of metadata repositories and metadata-driven statistical information systems. Collection of metadata includes the basic create, replace, update, and delete functions for any database. However, most organizational units responsible for surveys manage metadata in their own particular ways and the work necessary to put this information retroactively in a generally accessible database is overwhelming. This paper outlines the principles underlying good metadata management practice and provides examples.

Key Words: Co-ordination, Integration, Meta-information, Metadata Management, Standards, Statistical Information Systems, Surveys, Classification

1. INTRODUCTION

This paper builds on the ideas presented by Dippo and Sundgren (2000) and Gillman and Bargmeyer (2000) by describing the principles and practices of real world implementation. Section 2 outlines the requirements for metadata at various stages in the life cycle of statistical data, from survey design through collection, processing, dissemination, evaluation and back to design. The ability to link data and metadata more closely is stimulating the demand for new and more detailed information. Section 3 elaborates the metadata requirements for dissemination in more detail. Particular attention is paid to classification of statistical information to enable users to find, evaluate and access specific statistical data. Section 4 describes the methods by which metadata have traditionally been collected and maintained and the problems with these methods. In particular, the increasing demand for metadata is causing an increasing metadata burden and the need for new tools to assist survey management in automatic metadata creation. In Section 5 we propose some basic metadata management principles and procedures. Our concluding remarks are followed by the references from which we have drawn the material for this paper.

While the paper is of most directly relevant to the work of statistical agencies, it may also be of interest in the context of data archives (European model), polling firms, and universities that run surveys. Section 3, which deals with the issues of finding information resources, is of broad relevance to the work of librarians and information specialists.

2. METADATA CREATION AND USE DURING THE SURVEY CYCLE

2.1. Introductory Remarks

Before launching the discussion on metadata for survey design and dissemination, it is useful to consider the role of metadata in the use of statistics.

"Whereas the creators and the primary users of statistics might possess "undocumented" and informal knowledge, which will guide them in the analysis process, secondary users must rely on the amount of formal metadata that travels along with the data in order to exploit their full potential. For this reason it might be said that social science data are only made accessible through their metadata. Without human language description of the various elements, data resources will manifest themselves as more or less meaningless collections of numbers to the end-users. The metadata provide the bridges between the producers of data and their users and convey information that is essential for secondary analysis" (Ryssevik, 1999)

The creation of metadata can be a painstaking and demanding task. If statistical producers were required to produce metadata only for external users, these metadata would be seen as a major burden on time and resources.

Fortunately, metadata are created and required throughout the survey data life cycle and, with the right approach, much of the metadata data used in designing a survey will also satisfy the needs of external users.

The following paragraphs illustrate how metadata needs may be identified by outlining the creation and use of metadata in generic terms throughout the survey data life cycle. To facilitate the description, metadata are divided into five broad categories. The data life cycle is characterised as a sequence of four more or less sequential components: design and development; sample selection, collection and capture; editing, imputation, estimation, analysis and evaluation; and tabulation, dissemination and archiving. The data and metadata requirements for each component are summarised. With some modifications, this characterisation can also be applied to the acquisition and processing of data from an administrative source, or the merging and compilation of data from several different sources as in the national accounts.

2.2. Categories of Metadata

As there are many different types of metadata it is useful to group them into broad categories. In this paper they are classified into five groups, as follows:

- ***definitional metadata*** - describing statistical units, populations, classifications, data elements, standard questions and question modules, collection instruments, and statistical terminology;
- ***procedural metadata*** **-** describing the procedures by which data are collected, processed, disseminated and archived;
- ***operational metadata*** - arising from and summarising the results of implementing the procedures, including measures of respondent burden, response rates, edit failure rates, costs, and other quality and performance indicators.
- ***systems metadata*** - including locations, record layouts, database schemas, access paths used by programs;
- ***dataset metadata*** - a particular type of systems metadata comprising the minimal metadata required to describe, access, update and disseminate datasets, including title, textual description, data elements, data cell annotations and (optionally) population, source, and topics. Dataset metadata are categorised separately from other systems metadata because of the major role they play in dissemination and secondary analysis.

2.3. Survey Design and Development Metadata

Design and development commonly involves a (more or less comprehensive) redesign and redevelopment of an existing survey for a future survey cycle, rather than design and development from scratch, as entirely new surveys are relatively rare. However, as there is little difference, apart from one of scale, between design and redesign; the metadata requirements are essentially the same in either case.

The basic elements of the survey design and development process are to: establish the user requirements; define the target population(s), classifications, output data items and corresponding input data items; define the sampling frame, stratification and sample selection procedures; formulate the questions and reporting instructions, and select the collection instrument(s); define the data acquisition, capture, follow-up, and editing procedures; define the imputation, derivation, estimation, seasonal adjustment, analysis and evaluation procedures; and define the tabulation, deconfidentialising and dissemination procedures.

Data: Design and development do not entail any collection or processing of data per se, though they may involve manipulation of data from previous survey cycles, or related surveys, to obtain the metadata such as means and standard errors, required for sample design.

Metadata: The metadata for this survey component are primarily definitional or procedural. There are essentially two distinct modes of metadata use. First, there is access and analysis of the relevant metadata standards and of metadata for previous cycles of the same survey and related surveys. An important aspect of these metadata are supporting notes and feedback on problems previously encountered, solutions adopted and decisions made. Second, there is generation and recording of new metadata for use in subsequent components. The types of metadata required include:

- lists of populations/statistical units, classifications, and data items;
- statistical unit counts, sample design procedures, sampling intervals on random number line (for coordinated

surveys);

- forms/survey instruments, questions, instructions, and edit rules;
- collection, capture, follow-up and editing instructions and procedures;
- imputation rules, data item derivations, estimation formulae, seasonal adjustment formulae, analysis and evaluation procedures;
- tabulation and dissemination procedures, and 'deconfidentialising' rules

2.4. Sample Selection, Data Collection and Capture Metadata

The basic elements of this survey component are to: create the survey frame and select the sample; collect data by mail questionnaire/ interview/ electronic form; capture or transfer data to an input database; and follow up and resolve non-responses and input edit failures.

Data: Typically, an input database shell is generated from the survey frame. As raw data from respondents and from administrative or other sources are received they are recorded in an input database, usually in their original form. Subsequently, as these data are edited and supplemented, the corresponding updates are made to the database.

Metadata: Two categories of metadata activities are involved: first, definitional and procedural metadata generated by the design component are accessed and used; and second, systems and operational (results) metadata are recorded for transmission to subsequent components, and for performance analysis. The types of metadata required include:

- population(s) and classifications; frame, statistical unit counts, sampling parameters and sampling interval on random number line (for co-ordinated surveys); classification and contact data for the selected sample;
- form(s) or other collection instruments, survey procedures, and operational results, e.g., non-response rates;
- capture procedures, and results, e.g., percent captured and capture error rates;
- follow-up and input editing instructions and procedures, and results, e.g., follow up and error correction rates.

2.5. Editing, Imputation, Estimation, Analysis and Evaluation Metadata

The basic elements of these activities are to: edit the incoming object data to generate clean unit record data; impute for missing or inconsistent data to produce clean, complete microdata; aggregate, weight and derive data items, to generate macrodata; compute relative standard errors and other quality measures; compute seasonally adjusted and trend series, where appropriate; analyse output by comparison with data from previous survey cycles and other related surveys; and evaluate performance of design and operational procedures.

Data: Unit record data in the input database are transformed into micro datasets and, in most cases, subsequently aggregated to form macro datasets. (In certain classes of social surveys, output tables are generated directly from microdata.)

Metadata: The component involves access to definitional and procedural metadata from previous components, generation of dataset metadata for subsequent use by the tabulation and dissemination components, and recording of operational results for performance analysis. The types of metadata required include:

- edit and imputation rules, and corresponding operational metadata (error and imputation rates, etc);
- data item derivation rules:
- estimation formulae, seasonal adjustment and trend calculation formulae;
- relative standard errors;
- dataset metadata (name, descriptions, population, source, topics, data cell annotations, and data items);
- analysis and evaluation procedures, and the corresponding results.

2.6. Tabulation, Dissemination and Archiving Metadata

The basic elements of these activities are to: specify and produce output tables for regular publications and in response to special requests; 'deconfidentialize' macro and micro data outputs prior to release; and market and distribute outputs. In principle, the tabulation and dissemination systems are shared by surveys and may not be specific to a particular survey.

Data: Output tables are generated from macro and micro datasets, treated in accordance with confidentiality procedures, and distributed in a variety of formats and media.

Metadata: The component involves recording and use of dataset metadata to locate and retrieve datasets, retrieval of definitional, procedural and dataset metadata to accompany the data outputs, and recording of operational results for performance analysis. The types of metadata required include:

- definitional and procedural metadata (all types)
- dataset metadata (names, textual descriptions, populations, sources, topics, annotations, and data items);
- 'deconfidentialising' rules, and results of their application;
- tabulation and dissemination procedures, and results;
- marketing and distribution procedures, and results.

2.7. Scope of Metadata Elements

The scope of individual metadata elements may be broad, intermediate or narrow in the sense of being:

- shared across statistical agencies;
- shared across surveys or datasets within an agency;
- shared across survey cycles within a single survey program;
- for a single survey cycle or single dataset only.

Most definitional metadata, for example standard definitions of data elements such as "profit before taxes" or "period unemployed", have broad scope. They exist independently of surveys in the sense that they are not tied to any particular collection and may be used by several. Some definitional metadata, particularly derived output data items, may be specific to a single survey or even survey cycle. Procedural metadata, for example sampling or seasonal adjustment procedures, may be general and used by several surveys, or, like actual response and data item imputation rates, they may be specific to a survey, or questionnaire, or cycle. Operational and systems metadata are usually specific to a survey cycle or form. Dataset metadata elements, such as dataset name, footnotes and data cell annotation are specific to a single dataset. Others, like parametric (quantifiable) and classificatory (categorical) data items are values which may be shared by several datasets.

The scope of metadata determines, to a large extent, the utility of maintaining them in a repository that is broadly accessible.

2.8. Passive and Active Metadata

Another significant categorization of metadata elements concerns the nature of the role that they play in the collection, processing or dissemination of data. Metadata that determine the actions of automated survey and dissemination processes are referred to in this paper as "active metadata". They are sometimes called "prescriptive metadata. Those that simply document and inform about a survey process or dataset are referred to as "passive metadata" (or "descriptive metadata").

Evidently active metadata are vital to the survey or dissemination processes, and as such will certainly be stored and accessed as required. Passive metadata on the other hand are optional as far as automated data processing, dissemination or retrieval are concerned. Thus there is generally less concern that they are recorded at the time they can first be created, or that they are subsequently made readily available.

Statistics Netherlands has prepared a paper describing the potential use of active metadata tools in designing a survey. For more information, see Jean Pierre Kent (1999).

2.9. Metadata Models

Thus, metadata management implies the storage, manipulation and transfer of metadata through a sequence of collection, processing, dissemination and archiving. It involves a very broad range of metadata types, for each of which possibly different generation, storage, manipulation and retrieval facilities are required. In order to summarize

the requirements, it is convenient to identify a basic set of objects to which the metadata refer and to indicate their properties and relationships in a metadata context model. Figure 3 provides an example from the Australian Bureau of Statistics. For more details, see Colledge (1999) from which the following description was drawn.

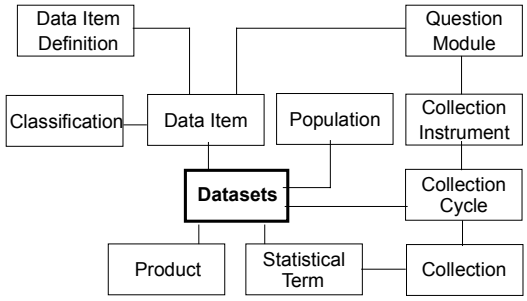


Figure 3. Warehouse High Level Data Model

At the core of the model, reflecting its output orientation, is the dataset entity. A dataset identifies a table of data together with corresponding row, column and cell annotations. It is linked to the population and data items described by the data, and to the products that draw on the data. It is also associated with one or more topics drawn from a prescribed list. Topics are used by search routines to facilitate identification of required datasets. The search routines may access synonyms and related terms recorded in a thesaurus in order to extend search power. The dataset is the principal object of data searches, being preferred to the data item entity, because data items never appear singly, they always appear in a context, and that context is embodied in a dataset.

A dataset is generated from one or more cycles of a collection (survey or administrative source). Associated with each collection cycle are descriptions of both the procedures it embodies and the operational metadata resulting from the cycle, for example response rates and cost. Also associated with a collection cycle is the collection instrument (or instruments), comprising question modules, which generate data items. Each data item is linked to a definition and other information relevant to its conception and use. In addition, a qualitative data item is linked to the classification that defines its value set. A glossary of statistical terms, linked to the thesaurus, is available to support descriptions, not only of datasets and collection cycles as indicated in the figure, but all other entities.

A model such as this, whilst being an abstraction, is a useful means of conveying the totality of metadata requirements.

3. METADATA FOR ACCCESS AND INTERPRETATION

3.1. Introductory Remarks

Up to this point in the paper, we have taken the orientation of a statistical producer going through the various steps of designing and conducting a survey. Once surveys reach the dissemination stage, it is reasonable to think about the metadata requirements of the users who may wish to find and access this information. The focus of such an examination shifts from an individual survey to a program of surveys. The challenge for a user, who may not be familiar with the details of these surveys, is to find the appropriate source of data to address his or her problem. The first step in a country with a decentralised statistical system like the United States may be to find out which statistical agency, archive, or other public or private body has the required information in its collection. Survey collections can be managed through the use of various classification and cataloguing schemes, which allow users to **find, evaluate and access** the data for use in their research and analysis (Boyko, 1998). This is a clear application of metadata in the sense that the user must examine surrogate information about the data rather than the data itself. The rest of this section will be devoted to the tools and approaches that may be used in finding the information and making the decision to use it for the challenge at hand.

3.2. Finding Statistical Information

Users start with questions or challenges for which they may choose to look for information. Going to a statistical agency, may or may not be their first instinct. Very few problems come labelled as to which type of information or data will shed light on them. As often as not, they may go to their library to look for this information or, in today's situation, may use the Internet. It should also be mentioned at this point that the term library also includes data library and data archive, each of which have staff that are extremely knowledgeable about data from statistical agencies and other sources. Whether the searching is done by the end user or by an intermediary and whether they are searching web sites or catalogues (which may be part of the web site or a standalone tool), well-chosen and structured bibliographic metadata are important elements of success. The metadata aspects of both approaches are dealt with below

3.2.1. Searching Web Sites

Most readers and users are familiar with the popular web search engines such as AltaVista, Lycos, Yahoo or Google. Research has shown that 25% of the users of Statistics Canada's web site reached it through a search engine. What may be less familiar, is the strategies that can be used for improving the probability that these search engines find sites that carry the survey information or products derived from the surveys. An ancillary issue is to make sure that the results rank as high as possible in the search results. The next challenge is finding the specific survey information on the site after the user has found the web site.

Strategies that will increase the probability that the statistical web site is found include **registering** it with the popular search engines and the use of the **meta** feature in **html**. Once the Uniform Resource Locator (URL) for a site is registered with a search engine, a program called a spider (sometimes called a 'bot' 'crawler' or 'worm') will be sent to the web site to retrieve information for indexing. How much information is taken and from where depends on the methodology of the search engine and the structure of the web site.

The use of the **meta** tag feature when developing web pages/sites, allows the creator to choose words that describe the content. This tag is not a 'required' one (i.e., the web page will work without it) but it increases the chances of a search engine having an appropriate entry in its index for the site. Useful tags for identifying statistical information would be **title, description** and **keywords.** The selection of keywords and the use of descriptive language in general should take into account the language that would likely be used those who will be searching for and on the site (although there are limits to how far one can go in this direction).

Once the user has reached the web site, the next steps are to browse or search its contents. Collections such as survey information or the products derived from them are often described in catalogues. Catalogues will be dealt with first and a few words about organizing a web site for **browsing** will be added later

3.2.2. Searching Catalogues

The first action of a data user or the data intermediary is to see if the required information is available locally. Most significant collections are best described in a systematic way, i.e., put in a catalogue. These are often referred to as OPACs or online public access catalogues. In today's world, OPACs may refer to or point to URLs, and websites as well as local electronic and hard copy information resources. There are a number of approaches that can be used to catalogue or describe a survey using standard headings. Some examples of metadata systems and tools used in this regard are shown below:

MARC (Machine Readable Cataloguing) The MARC formats are standards for the representation and communication of bibliographic and related information in machine-readable form and involves three elements: the record structure, the content designation, and the data content of the record. The MARC standard is widely used in the library community. It is of relevance to statistical and other data organizations as much of their traditional published information is likely catalogued in MARC format somewhere in their country. Statistics Canada catalogues all of its published products using this standard and is also in the process of cataloguing its data files and databases in MARC format. The MARC standard is particularly powerful when it is combined with Z39.50 compliant search engines.

Z39.50. The ANSI/NISO Z39.50 Search and Retrieval Protocol (also called ISO23950) is an open communications protocol standard for use in searching and retrieving electronic information over a network. One of the major advantages of using Z39.50 is that it enables uniform access to a large number of diverse and heterogeneous information sources. A client can thus provide users with a unified interface to both the query and the returned search results, regardless of where the information originated.

Dublin Core. The Dublin Core (named as a result of a meeting held in Dublin Ohio, USA) consists of 15 unstructured elements, which can be used to describe a variety of information resources. Its particular strength is that it is simpler than MARC but can still be used as an interdisciplinary record.

Warwick Framework. The Warwick Workshop was convened a year after the Dublin Core was developed to build on the Dublin results. It is intended to provide a more concrete and operationally useable formulation of the Dublin Core (and other meta-data schemes), in order to promote greater record interchange among content providers, content cataloguers and indexers, and automated resource discovery and description systems. It defines the containers into which Dublin Core content can stored. (Lamoze, 1996)

GILS (Government Information Locator System) GILS is also referred to as the Global Information Locator System due to the influence of the Global Change Program. It may not be global in practise but it is being used by the Canadian and US governments as a tool for the identification of government information resources. GILS was influenced by MARC and was designed for Z39.50 servers and clients. Its core element set (which is not reproduced here) is much larger than the Dublin Core. It goes into such areas as point of contact information, the provenance/source of the information, a number of administrative fields and fields for copyright and access constraints. It can be of relevance when a country's government decides that it wants to put a common face on the various web resources contained by different departments. Presumably, the same thing could happen using another standard. More information about GILS can be found at **http://gils.gc.ca**

3.2.3. Browsing Web Sites

Research indicates that only about half of the users of the World Wide Web (WWW) are comfortable or proficient in using search engines and Boolean logic for structuring their queries. A significant number of web users prefer to 'drill down' using the web's content structure. For this to be successful, the organization of the web content must be logical. One way of helping users 'drill' to the right spot on the site is to provide a subject index, which allows users to make successive choices on a topic list until they are at or near the information they are looking for.

There are many different ways in which metadata items can be classified to ease the difficulty of finding the right information when required. One such way is the classify them by subject. Subject lists created by analysing and organizing the key words pertaining to the content of a site or collections on contained on the site, can provide an intuitive entry point to browsing for data. One way of ensuring that this **indexing** is done in a consistent manner is to use a **thesaurus**. A thesaurus represents the relationships among the range of words pertinent to a subject area. It may have to be constructed or adapted from other organizations. A thesaurus includes such concepts as 'broader', 'narrower' and 'preferred' terms. Indexing the information items on a web site using such a tool can help users select an appropriate search word for a search engine or it can be used to organize content in a hierarchical manner as suggested above in relation to subject lists.

3.3. Evaluating Information

Let's assume our researcher has found a number of references that appear to meet the criteria set at the outset and that the main task is now to evaluate their relevance. This brings us to what could be referred to as a second level of metadata. If the metadata for **finding** data are referred to as **bibliographic metadata**, then the second level could be referred to as **statistical metadata**

At this stage, the user is interested in concepts, definitions, sources and methods. What universe is covered? What questions were the respondents asked? What definitions have been used? How have standard classifications been applied? What type of data or information has been found?

Before continuing this discussion, it may be useful to consider the different types of statistical metadata that could be encountered as a result of a search.

- **Summaries and analytical highlights** provided to the public when a survey or a study is first released to the public.
- **Journal and other articles** which describe the survey or study and present the finding
- **A statistical publication** with analysis and aggregate data tables.
- A statistical series or table in a **database** of aggregate data.
- **Public use microdata files (PUMFs)-** Files that have been anonymized and can be released to users without identifying respondents.
- **The survey master file** (which will contain confidential information and may thus be of relevance only within a secure environment such as a statistical organization or a research data centre)
- **Statistical infrastructure** such as classification systems.

The search engine points to the (bibliographic) metadata and the bibliographic metadata points to the statistical metadata which describes the resource in sufficient detail for the user to decide whether it is relevant to their work. This type of linkage is straightforward where there is a clear link between the survey output and the various parts of the survey description. For PUMFs, it is not unreasonable to have the entire codebook on the website. Finding references to articles based on specific data sets requires that sufficient referencing and indexing have been done. Documenting national accounts series is a different kind of challenge as their content is derived from the manipulation and aggregation of survey aggregates. However, even if the metadata for different types of statistics vary somewhat, the principles described above still hold.

There is no standard set of fields for these type of metadata in the same way that there is for bibliographic metadata. There are, however, some attempts being made to specify such standards. These will be mentioned below.

3.4. Accessing and Using Information

Having found a data source and decided that it is relevant, users want access. If the user is in a library, he/she hopes that the information resource is available locally or can be obtained quickly. In a 'point and click world', users expect to be able to access information directly from websites and databases. One way in which this can be done is to have linkages from the metadata used to find and evaluate data to the actual data. Ideally, the data that are linked to should be in a format that allows it to be used in a variety of analytical software. While there is not single ideal format, there are a number of widely used formats, eg comma or space delimited ASCII files. Proprietary formats should be avoided, but if they are be employed, the user should also be given other more generic choices.

The users of public use microdata files absolutely depend on the codebooks that accompany these files. One difficulty that they often encounter is an underdeveloped standard for the production of these codebooks and a lack of adherence to the standards that do exist. In addition, there is often a lack of appropriate formats to enable the data to be used directly without considerable effort on the part of the user. Since software such SAS, SPSS and STATA are widely used in statistical analysis, users generally prefer to have these code sets available with the data files.

3.5. Archiving and Preservation

Most studies and surveys are produced and used within a fairly short period of time (say years or even decades). But they also may have considerable value over time (say 50 to 100 years, or even indefinitely) and it is for this reason that in many countries, the disposition and/or the long term archival management of these data is the responsibility lies with an archiving authority. This requires data formats and data documentation that are sufficiently generic and understandable to allow future generations of researchers and statisticians to be able to access and understand this information in the future. Formats and storage media that permit the migration of the data and the metadata over time and across generations of media are crucial for long preservation.

4. PROBLEMS IN CURRENT METADATA MANAGEMENT PRACTICES

The current problems with metadata management affect both the users and producers of data. Metadata are often incomplete, inconsistent and not shared; they may not follow standards but rather represent what was possible under the circumstances. At the same time, improvements to metadata may result in tasks (creating and collecting metadata) which can be very time consuming and expensive for the survey producers.

Incomplete metadata: Insufficient metadata for search engines to find the information eg, no index terms; insufficient information at the question, concept and variable level for researchers to know whether the source meets their needs; insufficient formatting for the data to 'slide' easily into an analysis software. The standard set of metadata fields desired for a statistical resource is not well defined. Incomplete metadata are of limited help in designing or redesigning surveys. While there are standards for bibliographic metadata, their use for documenting statistical files is at an early stage. Similarly, organizing and presenting statistical data on the web is still developing as a practice.

Inconsistent metadata: Surveys brought together into a collection vary in terms of content detail and format. This makes it difficult to organize for browsing and searching. The lack of adherence to standards makes the data and metadata more difficult to use where they are managed as part of a collection of surveys. Consistency is a prerequisite to making data and metadata interoperable with software.

Unshared*:* If metadata are not shared, then opportunities of efficiency and effectiveness are lost. This is true both for survey design and operations as well as for dissemination. While there may be some costs involved in organizing the sharing of metadata, gains are also possible. The development of tools for sharing metadata can result in both improved consistency and more complete metadata as noted by Gillman and Bargmeyer (2000). Publishing metadata on web servers is a good way to share them across applications and provide visibility for them; but this also makes inconsistencies more obvious.

Some of the problems in current metadata management practices arise from the ways in which survey programs tend to have grown over time. Typically, surveys have been developed independently of one another, in response to specific separate needs. Thus an agency's survey program often comprises a set of semi-autonomous "stovepipe" survey operations, sharing few, if any, procedures or software. This affects the use of metadata not only for the design and production functions but also for dissemination activities.

In this sort of compartmentalized environment, metadata are rarely shared. Active metadata are, from necessity recorded, but are often inaccessible other than to the programs for which they were specifically created. Passive metadata are often not recorded at all, or are recorded long after the time of their creation. There is thus a lack, duplication or inconsistency of metadata across surveys. The traditional approach to building metadata content has been to focus on one survey at a time and present this information to users via summary publications and as part the paper code books that accompanied machine readable data sets.

Costly and time consuming*:* In constructing cross-survey metadata repositories, the metadata elements are often collected from survey managers using paper questionnaires, long after the events to which the metadata refer. The resulting information is captured again, catalogued by experienced cataloguers/indexers and disseminated on various media. Such one-off, paper-based metadata are costly to produce, difficult to maintain, and of uneven quality across the agency. This is another argument for organizing corporate repositories.

5. GUIDING PRINCIPLES FOR METADATA MANAGEMENT

5.1. The Ideal Metadata World

When data collections and elements are described according to metadata standards, tools can be created to share and deliver them in useful ways. In this way, users at all levels are empowered to conduct their own analyses and research, decision makers and practitioners become more accountable for the ensuing outcomes, and billions of dollars in investments in data are leveraged as never before possible. In an ideal world, all the metadata for a statistical agency would be automatically recorded at the moment of their creation during the survey or data

dissemination process, and would be available subsequently to any user or program that had need of them. Survey design tools and operating procedures would have metadata creation imbedded within them. Creation would be in accordance with appropriate international, national or agency standards for the different types of metadata. For example, questionnaire designers would be able to choose from among the set of concepts and definitions that were recommended for use throughout their agency so that results from different surveys could be readily compared and integrated. The codes assigned to variables would be standardized, making it easier to create and use the files. By locking datasets and the corresponding metadata together, survey outputs would be self-describing, thus supporting their interpretation and use. Dataset and associated metadata formats would follow internationally accepted standards to facilitate searching, accessing and use. They would be readable and browsable by common statistical software. Finally, all data and metadata would be rendered into generic, non-proprietary formats for long term preservation and migration to new platforms.

In such a scenario, metadata for multiple use would be held in a metadata repository or "registry" as Gillman and Bargmeyer (2000) prefer to call it. The key features of such a repository are:

- that metadata it holds are created in accordance with well defined rules regarding their attributes - it is in this sense that they are "registered";
- that it incorporates international, national and agency standards to facilitate multiple users and uses;
- that it includes all the various types of metadata and their linkages;
- that it is broadly accessible, typically web-enabled.

The conversion of metadata from one schema to another would be handled through the use of SGML (Standard General Markup Language) or XML (Extensible Markup Language) tags and DTD's (Document Type Definition). XML promises to be the new language of the WWW, replacing HTML (Hypertext Markup Language). XML is a simplified version of SGML, which is a language for describing languages. (HTML is an instance of SGML). The power of XML will lie in the fact that it will be usable to design tags (labels) that portray the subject matter of an area to which it is applied. For instance, in the case of the Data Documentation Initiative (DDI) which will be described below, an XML tag will define a variable so that when a statistical software reads it, it will know what to do with it. XML will be readable by browsers in the same way as HTML whereas SGML was simply too complicated for a browser to handle.

The Resource Description Framework (RDF) is a sister standard to XML and will be used to create metadata for the WWW. Jon Bosak and Tim Bray (1999) - both members of the World Wide Web Consortium (W3C) which is shaping the future of the WWW - describe XML and RDF in lay terms as follows:

"Resource Description Framework (RDF) is a foundation for processing metadata; it provides interoperability between applications that exchange machine-understandable information on the Web. RDF emphasizes facilities to enable automated processing of Web resources. RDF can be used in a variety of application areas; for example: in resource discovery to provide better search engine capabilities, in cataloguing (sic) for describing the content and content relationships available at a particular Web site, page, or digital library, by intelligent software agents to facilitate knowledge sharing and exchange, in content rating, in describing collections of pages that represent a single logical "document", for describing intellectual property rights of Web pages, and for expressing the privacy preferences of a user as well as the privacy policies of a Web site. RDF with digital signatures will be key to building the "Web of Trust" for electronic commerce, collaboration, and other applications."

Developments such as those noted above will make it easier to link the different types of metadata together. The consistent and persistent use of index terms taken from a thesaurus and attached to the information resources will make them easier to find and will enable cross-references to published analysis and research based on the data. Thus one of the options open to a researcher who finds a data set will be to link see who has published what using the data and may even see references to challenges faced in using the data.

5.2 Metadata Quality

Over the past decade, as a result of the quality management drive, in particular the recognition of the user as the ultimate arbiter, the definition of quality applied to statistical data has broadened significantly. In a departure from

the earlier narrow focus on accuracy, Fecso (1989) introduced four major quality components: relevance, accuracy, timeliness, and cost. Subsequently, these have been supplemented with more components to deal with dissemination -- visibility and accessibility -- and reflecting the capacity to use data from different sources -- comparability and coherence. Thus data quality is now viewed in terms of these various components.

Given that metadata describing data are themselves data, their quality can be assessed in accordance with these same components along the following lines.

- **Relevance:** Do the metadata match the users' needs? Are they useful? Would other metadata be more useful?
- **Accuracy:** Do the metadata convey an accurate impression of the quality of the data or survey process or data elements they describe?
- **Timeliness:** Are the metadata available when they are needed, e.g. at the time of analyzing the data, or redesigning the survey?
- **Cost:** Can the cost of their production be justified in terms of user benefit? Would other possible uses of the same resources be more beneficial overall?
- **Visibility and Accessibility:** Are the metadata readily visible to users? Accessible to users?
- **Coherence:** Are the metadata coherent over survey cycles? Do they reflect real changes? Do they include 'artefactual' changes resulting from the way they have created?
- **Comparability**: Are the metadata readily comparable across datasets, surveys, agencies?

5.3 Steps towards Better Metadata

The following provides some guidelines for moving towards this ideal metadata world.

- Identify the metadata requirements and develop a rudimentary metadata model along the lines previously outlined in Section 2.
- Identify metadata standards appropriate for each of the various types of metadata. Use internationally recognised standards where available, for example United Nations (1995), also see Byfuglien (1998), and Gillman and Bargmeyer (2000). In the absence of international standards use national standards, and finally agency standards for everything else, developing them from scratch if need be. Recognize that there may be more than one standard for creating statistical and bibliographic metadata.
- Develop a metadata registry to house all the categories of metadata. This could comprise a single database or more likely, taking advantage of repositories that already exist, it might consist of a set of connected databases. The important thing is that the registry can be viewed as a whole, i.e., that it is conceptually complete and the metadata are linked together as they should be.
- The initial load of the registry can be achieved by assembling all available information from existing sources and passed to the survey management unit for verification. Updates to the base information can be supplied online and gradually improved if necessary. Loading of current survey information should be the first priority but in many cases it may also make sense to load some of the retrospective survey information as well. Having survey metadata available on a public website can motivate survey managers to ensure that the information is as up to date and accurate as possible. Nevertheless, the initial creation of the registry will be rather time consuming and resource intensive.
- Develop general-purpose tools for loading and updating metadata in the registry. Recognition of the potential of electronic collection of metadata is leading to the development of electronic tools that can assist in this process. A general approach to electronic collection of metadata includes the use of a series of electronic screens deployable over a network to be filled in by the various specialists involved in survey design and operation.
- Ensure that, for each new or redesigned survey, metadata creation is built into the systems and procedures and that it occurs at the time the metadata are first available, not retrospectively.
- Exploit technologies such as XML to facilitate the conversion and interoperability of data and metadata with statistical softwares.
- For existing surveys, ensure that the persons most closely associated with particular aspects of the survey, eg, design, collection, processing, are assigned the tasks of passive metadata creation. The use of 'pick lists' rather than free format metadata makes their task easier and the results more consistent across surveys. For textual metadata, ensure cut and paste facilities are readily available so that survey staff can readily extract from survey records and paste into the registry.

- So far as possible, make metadata active, if need be by the artifice of requiring metadata completion as a condition for proceeding further in collection processing or dissemination.
- Develop and promulgate an Agency wide metadata management policy to support the recording and quality of metadata, including metadata quality measures.
- Develop and supply key words and index terms that can facilitate broad based resource discovery on the web or in catalogues.

6. CONCUDING REMARKS

Statisticians have used the term metadata for some time to refer to information and data used to describe data. The emergence of the WWW as a network of interrelated information nodes through which one must navigate has propelled the use of this term into a much broader context. It is also evident that professions such as librarians have been working with metadata to carry out the tasks of organizing access to information for a long time. In the context of the work of statistical and data organizations, it should be noted that there are similarities between the type of metadata used to design and process surveys and the metadata that can be used to facilitate data dissemination and access. We are starting to see the emergence of new data design and processing environments, which have the potential to feed downstream activities such as dissemination, access and preservation. This in turn leads to a demand for the use of integrated metadata systems.

In terms of implementation, current practices are lagging a long way behind accepted principles. Even at the agencies cited in the previous section as providing good examples, much metadata are not recorded at all or are inaccessible. The first step in moving to a more ideal situation is to realize the need for metadata management and to introduce the basic tools. A metadata registry is a good starting point as it can be used to achieve a broad range of objectives. However, implementation of such registries for existing and new surveys may still be a challenge.

In the case of new surveys there are no good reasons for not implementing good metadata design practices. However, many of the major current survey systems at statistical agencies were designed and built long before the focus on metadata. They thus generate automatically only the metadata absolutely vital to further processing or dissemination of the data, i.e., in the terminology of this paper the "dataset metadata". Given that complete re-engineering of these systems cannot be achieved overnight and must be handled incrementally, metadata management must be largely retrofitted. In the short run, progress can be made by utilizing and joining up with the metadata that may have created by others to facilitate access and use of data. The quest for web compatibility leads to tools such as SGML and more recently, XML which are being used by a broad range of projects and subject areas to facilitate publishing and information exchange. Projects such as the ICPSR's DDI are using these tools and a panel of experts to develop a standard way of describing statistical surveys and studies. The main output from this initiative will be a DTD, which can be used by others working with data files. Use of the DDI standard by projects such as NESSTAR and the Harvard Virtual Data Centre project would suggest that it will emerge as standard that will be applicable to the work of statistical agencies. While its main focus thus far has been at the survey level (micro data), phase II of the project will be concentrating on aggregate data to a greater extent.

Creating good metadata may in fact be difficult and demanding work. However, with the growing expectations of users and the potential benefits of using corporate metadata for designing and implementing surveys, there is also the potential for gains in efficiency, consistency accuracy and useability. Good metadata may be able to 'pay' for itself in terms of increased data analysis and knowledge creation.

REFERENCES

- Bosak, J. and Bray, T. (1999) Scientific American, May 1999, **http://www.sciam.com/1999/0599issue/0599bosak.html**
- Boyko, E.S. (1998), Statistics Canada, "The Evolution of Metadata at Statistics Canada:: An Integrative Approach" Contributed paper for the 1998 meeting of the METIS working Group, UNECE, Geneva, Switzerland. Available at **http://www.unece.org/stats/documents/1998/02/metis/13.e.html**
- Byfuglien, J. (1998) Statistics Norway, "Standards for Meta-data on the Internet", Paper prepared by for the 46the meeting of the Conference European Statisticians, Paper CES/1998/32) May 18-20, 1998, Paris France. Available at **http://www.unece.org/stats/documents/ces/1998/32.e.html**
- Colledge, M. J. (1999), "Statistical Integration Through Metadata Management," *International Statistical Review*, **67**, 1 pp. 79-98
- Dippo, C and Sundgren, B. (2000) "The Role of Metadata In Statistics", International Conference on Establishment Surveys-II, June 17-21, 2000, Buffalo, New York.
- Fecso (1999), "What is Survey Quality: Back to the Future", pp88-96, Proceedings of the Section on Survey Research Methods, Annual Statistical Meetings, American Statistical Association, Washington, DC.
- Gillman and Bargmeyer (2000) "Metadata Standards and Metadata Registries: An Overview, ", International Conference on Establishment Surveys-II, June 17-21, 2000, Buffalo, New York.
- Ryssevik, J. "Providing Global Access to Distributed Data Through Metadata Standardisation-The Parallel Stories of NESSTAR and the DDI. Contributed paper for the 1999 meeting of the METIS working Group, UNECE, Geneva, Switzerland.
- Statistics Norway (1998), "Guidelines for Metadata on the Internet," paper presented at UN/ECE Conference of European Statisticians, June 1997.
- United Nations (1995), "Guidelines for the Modelling of Statistical Data and Metadata," Conference of European Statisticians Methodological Material, United Nations Statistical Commission and Economic Commission for Europe.