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; Sundgren (1994).

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 systems). 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 adminitstrative 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 also 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 as published by May, 2000 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’re 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’re 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 Census Bureau 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. Otherwize 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 of so-called process data, is 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 the data collection processes, the data collectors, the data processors, and the 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 such 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 (in the sense of the infological equation; see above) 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 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, an illustrative example and 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 producers 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, 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/country, a time period, or branch of industry. 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 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 finds 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.

An important recommendation from this research was the need to rethink the interface to the BLS website (which reflects BLS’ program-oriented organization) so that it would better meet the needs of users with diverse expertise and needs. Based on these results, Marchionini (1998) proceeded to design and test alternative interface designs. The iteratively-developed designs were 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, and a particular convention provided in an informative annex is being taken into account in this research. 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. The pilot work involved generating short variable names based on abstracting language from a survey question and its valid responses. 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 or grammaticality had little influence on comprehension. On the other hand, even this small test indicates that it may be more difficult to find any "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 re-designed 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 involving Carol Hert and staff from both BLS and Census 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.[2]

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 non-redundant 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 Bengt Rosén (professor of statistics) and Bo Sundgren (professor of informatics); see 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, 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) products 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; Statistics Sweden (1995) and 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

[2] To our best knowledge, the originator of the slogan is Wouter Keller, Statistics Netherlands.

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 withinthe 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. It's breadth of meaning is such that the metadata producer must address it's 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.

7 References

Deming, W. E. (1982), *Quality, Productivity, and Competitive Position*, Cambridge, MA: Massachusetts Insitute 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: JohnWiley & 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 (1994), *"Statistical Metadata - A Tutorial"*, Invited paper for the Compstat conference in Vienna.

Sundgren, Bo (1995), *"Making Statistical Data More Available"*. 50th session of the International Statistical Institute (ISI), Beijing 1995. Also available from Statistics Sweden: R&D Report 1995:6.

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.