PROPOSAL FOR A STATISTICAL METADATA STANDARD

William P. LaPlant, Jr., Gregory J. Lestina, Jr., Daniel W. Gillman, Martin V. Appel
U.S. Bureau of the Census

ABSTRACT

A statistical metadata content standard is being developed by the Bureau of the Census. The "Survey Design and Statistical Methodology Metadata" standard is an inclusive set of statistical metadata concepts which characterizes all aspects of survey design, processing, analysis, and data sets. This standard's development is a key step in designing a unified Census Bureau metadata repository for survey and census activities.

The goals for this effort are: 1) to develop agreement within the Census Bureau and among our sponsors on the set of metadata necessary to accurately describe our statistical data products, and then to seek advice and concurrence from other Federal users and the statistical community at large; 2) to provide a common thesaurus that can be used as the basis for the content of metadata tags (identification tags) within other standards such as the Spatial Data Transfer Standard [FIPS-173], the Government Information Locator Services Standard [FIPS-192], and proposed and existing (i.e., WAIS) standards for automated indexing on the Internet; 3) to serve as one of the sources of conceptual entities for the modeling needed to develop usable metadata repositories.

The proposed standard assumes the existence of the draft "Cultural and Demographic Data Metadata" (CDDM) standard [FGDC/SCDD-95]. In particular, the following components of the CDDM are used by explicit reference: "Themes," "Geographic Information," and "Temporal Information." Elements of the "Identification," "Source Information" and "Data Quality" components are incorporated, with modifications, throughout the proposal.

This paper describes the above items and provides the location on the Internet where those who are interested can review the standard and provide comments.

KEYWORDS
Survey design, statistical methodology, analysis, data sets, metadata repository, GILS

1. Introduction

The “Standard for Survey Design and Statistical Methodology Metadata” (SDSM) is a statistical metadata content standard. It is being developed at the U. S. Bureau of the Census (BOC) in support of various Census Reinvention Laboratory initiatives and other efforts[a]. These, in turn, are intended to support the mission of the Bureau:

To be the preeminent collector and provider
of timely, relevant, and quality data about
the people and economy
of the United States of America.

We intend that the SDSM provide a description of the information or documentation about statistical data. It is intended to be a comprehensive, hierarchical thesaurus of terms, an outline of all the concepts contained in *any* documentation about the design, processing, analysis or data dissemination of surveys or censuses. Users will not necessarily have access to every type of documentation listed. Access might be restricted for several reasons: sophistication of users; confidentiality; lack of relevance or existence; etc.

The Standard provides a mechanism for comparing and linking among statistical metamodels, obtaining consensus on statistical concepts independent of how those concepts are used.

This paper discusses:

- what statistical metadata is,
- the purpose of the standard,
- the relationship between this standard and other standards,
- who needs statistical metadata,
- goals for the project,
- what the standard contains,
- how the standard will be used, and
- the development schedule.

2. What is Statistical Metadata?

Statistical Metadata is descriptive information or documentation about 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.

Statistical data consists of the following:

Microdata — data on the characteristics of units of a population, such as individuals, households or establishments, collected by a census, survey, or experiment.

Macrodata — data derived from microdata by statistics on groups or aggregates, such as counts, means, or frequencies.

Metadata — data which describe the microdata, macrodata or other metadata.

The extensive nature of statistical metadata lends itself to categorization into three components or levels:

> Systems — the information about the physical characteristics of the application's data set(s), such as location, record layout, database schemas, media, size, etc;
>
> Applications — the descriptive information about the application's products and processes, such as sample designs, questionnaires, software, variable definitions, edit specifications, etc;
>
> Administrative — the management information, such as budgets, costs, schedules, etc.

These components, "Systems, Applications, and Administrative", help to differentiate the sources and uses of statistical metadata.

3. Project Goals

a. Develop Agreement Among Survey Developers and Users on Terms

It is our intention to involve in the development of this standard, personnel knowledgeable in all aspects of survey design, execution, analysis, and dissemination. Final coordination of the standard will include users of our data and the broader community of statisticians, demographers, and economists.

b. Define Terminology for Describing Metadata Content

We intend that the standard provide developers and users of statistical products with a common vocabulary for describing the design processing, analysis, and data sets for censuses and surveys. Broad agreement on the meaning and organization of these concepts will provide the basis for improved communication among the producers and users of economic and demographic statistical data sets. We recognize that it may not be possible for all to agree on every definition, so initially, many terms may have multiple definitions.

The SDSM also will serve as a glossary of statistical metadata concepts. Each term in the standard is defined. The terms are presented in the order in which they are likely to be encountered as surveys are designed and implemented; and as the results are analyzed, reported, stored and disseminated. Thus, any given term may have several different definitions, depending on the context of its use.

c. Develop a Metadata Repository

A metadata repository is a system that allows for the query, editing, and managing of metadata. Such a system provides a mechanism for looking up information about our statistical products as well as their design, development, and analysis.

4. Purpose of the SDSM Metadata Standard

a. Define Entities for Metadata Models

We plan to develop a Statistical Metadata Repository Database. The repository will ease the job of survey designers by ensuring the previous design documentation is available. The job of managers and planners will be simplified by making available status, personnel, logistics, and financial information in a uniform way. System designers and implementors will have access to component designs and code of earlier related systems. Each of these uses and others may need to be supported by separate views of the same entities.

b. Support Related Standards

There are a number of standards which can make use of information provided in populating a metadata repository. These include data interchange standards such as the Spatial Data Transfer Standard [FIPS-173], and the US indexing system service standard, ANSI/NISO Z39.50, and its profiles for automated indexing on the Internet, the Government Information Locator Services (GILS) [FIPS-192] and proposed and existing Wide-Area Information Services (WAIS).

5. Relationship to other Standards

The SDSM is intended to be used with the “Standard for Cultural and Demographic Data Metadata [FGDC/SCDD-95]” (CDDM) Draft. The CDDM, is in turn, mapped to the metadata portions of the “Spatial Data Transfer Standard [FIPS-173]” (SDTS) and supports providing metadata for the “Government Information Locator Service [FIPS-192]” (GILS). The SDSM assumes the existence of these other standards which define additional, related, metadata. The thematic content of a data file is provided as specified by the CDDM while the physical layout is provided by either an SDTS mapped to a “Data Descriptive File for Information Interchange [FIPS-123]” (DDF) specification or by a GILS specification.

The Spatial Data Transfer Standard (SDTS) is designed to assist in moving the contents of the Geographic Information Systems (GIS) databases between two dissimilar GIS servers or to exchange data between systems that have at least some capabilities for generating, analyzing, storing, or displaying geographic data[b].

The CDDM standard includes a mapping to the metadata components of the SDTS [FIPS-173], and follows the syntax and semantic rules specified in Part 1 of the SDTS. It is intended to be used as part of the metadata component of both GILS and SDTS. Upon completion, this standard will likely be issued as a FIPS.

The Government Information Locator Service is a decentralized collection of servers and associated information services. These servers and services will be used by the public, either directly or through intermediaries, to find public information throughout the Federal government. GILS servers will support search and retrieval by accepting a search query and returning a result set or diagnostic messages. Development of GILS by Federal Agencies is mandated by the “Paperwork Reduction Act of 1995.”

6. Who will use the Survey Design and Statistical Methodology Metadata (SDSM) Standard?

The metadata identified by the SDSM is intended to provide data users, designers, analysts, and others with a complete set of information about the data of interest.

Data users could ask:

- What information is available?
- What format is it?
- What is the definition of the data item?
- What is the quality of the data?
- How accurate is the data?
- How do I get the data?

Survey designers and analysts might ask:

- How was the sample designed?
- What were the editing rules?
- How well did our sampling strategy work?
- What was the quality of the data?
- What was the response rate?
- What processing methods were used?

To summarize, statistical metadata is intended to provide descriptive information or documentation about statistical data. The remainder of the paper will describe the standard, discuss our goals in developing the SDSM, describe the sources for the terms used, and how the completed standard can be used.

7. The SDSM Metadata Standard

The SDSM standard is a list of statistical metadata terms and their meaning.

a. Chapters

The SDSM is a "*content* standard." This means that it defines what metadata items about statistical surveys and censuses contain. The SDSM does not specify the physical format of the content, the services to be provided, or the syntax to be used. Each is the subject of other related standards mentioned elsewhere. The SDSM also supports labeling of the content of metadata either by tags included in the metadata itself, or by indexing included in various systems implementing mechanisms or tools.

The standard is designed to help the contributors and users of the metadata answer "Who," "Why," "What," "When," "Where" and "How" for Surveys, Systems, and Products.

The metadata items of the SDSM standard are organized into "chapters." Each chapter represents a logical set of metadata. The inclusion rule[1], and a definition is provided for each metadata data element. The initial, defining entry for each of the eight chapters are shown below as they appear in the standard.

0. **Identification** *(mandatory)*. This chapter contains the minimal set of mandatory metadata items and is applicable to the entire set of metadata. This section contains identifying information and any documentation developed during the conceptualization phase of survey planning.

[1] Shown in *italics*. For example, *(mandatory)* or *(optional).*

1. **Content** *(optional)*. This chapter contains information about the nature of the data that is the subject of the survey, i.e., the universe of interest and the specific data items to be gathered. Contains definitions, data standardization rules, and coding information.

2. **Planning** *(optional)*. Documentation related to the project **planning** for all phases of survey work. This includes documentation related to budgeting, staffing, and training.

3. **Design** *(optional)*. This chapter includes information on the development of the universe and frame; sampling strategies; the design of the "measurement instrument" (questionnaire or equivalent); the construction of the "observation register" including the check-in, check-out mechanism; and how non-response will be handled.

4. **Implementation** *(optional)*. This chapter includes documentation related to **implementation** of the survey, including: interviewer procedures, guidelines and training materials; distribution and collection of forms or other measurement instruments; execution of the "observation register," i.e., check-in, check-out and enumerator diaries; field edits and verification; follow-up procedures, training and tracking; sampling mechanism for follow-up and quality assurance; data preparation procedures and training; and mechanisms for creating and maintaining records on the process.

5. **Analysis** *(optional)*. Documentation related to all statistical processes used to analyze the survey results or those used for displaying or presenting the resultant information.

6. **Data_Processing** *(optional)*. (Computer Systems) Documentation of all computer processes needed to support survey activities or processes.

7. **Data** *(optional)*. Documentation concerning all data sets retained related to the survey, and, possibly, the data itself.

b. Metadata Data Elements

Each section consists of an outline of concepts. Each entry in the outline is a metadata data element. Any of these metadata data elements may be used to identify specific instances of metadata. The metadata itself may be a complete **Citation**, an **Abstract** of the information, the **Metadata** itself, or a description of how the information may be obtained electronicly. This last description must be provided as a **Uniform Resource Locators** (URL) [RFC-1738]. All of the above, **Citation**, **Abstract**, **Metadata**, and **URL**, may not apply to some metadata data elements. These cases will be obvious in the context of the formal definition provided in the "Data Element Outline."

Formal definition of how the metadata itself may be provided:

Any metadata data element of type "text" and without a constrained domain may be considered a compound metadata data element with any of the data elements below it in the hierarchy or any of the following additional data elements, in any combination forming the complete metadata data element definition:

- **Abstract**: A brief description or extract of the applicable text of the document or object.

- **Citation**: The full, formal citation used to reference the object.

- **URL**: A pointer to the referenced electronic document or object and its associated access service.[2]

- **Metadata Composite**: The **Metadata Composite** may contain the full content of the document or other 'object,' and may contain another **Metadata Composite** (be recursively defined). The nature of the content of a document or other object may be limited by implementation constraints. What an object is, is context dependent. This element actually contains the metadata. Generally this component will be some form of text. This definition is recursive to support complex data objects for electronic storage of documents such as Presentation Definition Format (PDF) or Office Document Architecture (ODA). Each of these supports more complex data objects containing multiple data types such as typographically formatted text, graphics, audio and spread-sheets formats.

The metadata elements that comprise a section may participate in describing the metadata data element above it in the hierarchy. Additionally, each metadata data element may be part of a compound descriptor; it and all the metadata data elements below it in the hierarchy form the complete description.

8. The Table of Contents View

One way to view the proposed metadata standard, is to imagine that all the metadata at the Census Bureau is contained in a book.

a. The Analogy

The SDSM is the table of contents for that book. Like a table of contents, the SDSM provides a map to the contents of the book. The TOC provides a way for "readers" to quickly go to areas of interest.

In use with survey documentation tools[3], the proposed metadata standard will be seen as a table of contents by the user (the survey designer, subject matter analyst, etc). If the user is working with existing documentation, the tools will assist in organizing and annotating that documentation, producing an "automatic table of contents" so that various documentation components would be accessible by other users in a uniform manner. If the user is designing a new survey, this tool will provide a ready-made structure for developing the required documentation. This will ensure that the various aspects of survey design and analysis are addressed, or at least that an explicit decision is made to defer addressing them.

[2] The URL is the Client/Server access pointer as specified in the Hypertex Mark-up Language (HTML) for use in the Hypertext Transfer Protocol (HTTP), as released by the HTTP Working Group of the Internet World Wide Web (WWW).

[3] This tool is being developed.

The TOC hierarchy has allowed us to further develop high level conceptual models of existing systems and show how they will interface with the proposed metadata repository. For example, the Integrated Processing System (IPS) will not store metadata but will link to the metadata repository to get the location of available metadata. The Standard Economic Processing System (StEPS)[4] will need a separate data element registry for assigning definitions to elements in their repository so that information can be standardized across different repositories. We are currently developing the conceptual and logical models for the standard metadata repository. The TOC is being used as the point of reference for these models.

b. Our On-line Table of Content (TOC)

The On-line TOC was developed to help reviewers of the standard. The on-line TOC presents, in an easily accessible way, on the World Wide Web. The TOC is a combination of HTML pages and HTTP CGI[5] scripts written in Perl[6]. The CGI scripts are used for displaying, navigating, and allowing users to enter comments about various elements of the standard. The on-line TOC provides an easy method for users to become familiar with the SDSM Standard and allows users to enter their questions or comments on any of the elements in the standard.

The on-line TOC is hierarchical in design and displays each section of the standard on a separate Web page. From a page, a user can select the next lower level in the hierarchy, or go to any page in the TOC by entering the section number in the dialogue box at the end of the page. Each page of the TOC displays all the sub-sections under the parent section in the hierarchy along with their section description. Near the top of each page of the TOC, there is the hierarchical path of sections leading to the current section level. A user can select any level superior to the current section by double-clicking on the section description. Each page also has a selection for making comments about the page. When a user wants to make comments, a new web page is displayed and the user enters the comments into a dialogue box.

In addition to being able to select a section in the TOC, each section has a button for displaying the definitions. By double-clicking on the “Definition” button, the definition for that section is displayed on a new page. After the definition, there is a dialogue box available for user comments. Access to the on-line TOC is available on <http://www.census.gov/ftp/pub/std/TOC.html>.

[4] StEPS is being developed by the Economic Planning and Coordination Division (EPCD) of the Bureau of the Census.

[5] Common Gateway Interface. This is a mechanism that provides WWW providers with the ability to provide WWW browsers with HTML from a *program* running on the server as opposed to a precoded HTML file. GCI scripts can provide a dynamic 2-way interface that is tailored to each user's unique request.

[6] A interpreted programming language.

9. Schedule

The following is a schedule[7] for the events remaining in our plan for the development of this standard. The plan is based on the experience of several of the authors with the standards development process both within and external to the BOC.

- ○ Now — Initial Draft for Informal Comments

  — (Open to all)

- ○ May — Census Colloquium on Standard and Policy

- ○ June — Census Standards Process Begins

- ○ Oct 1

  — Census Approves Statistical Metadata Standard

  — FIPS Coordination Process Begins

10. Conclusion

The SDSM Standard will be used to augment other standards and as the common view used by different metadata tools. For GILS, the SDSM may provide indexing terms which a GILS service will associate with a pointer to the metadata itself. For SDTS, the SDSM may provide a registered metadata identifier that serves as a sentinel for an extended set of metadata included in the SDTS encoded dataset.

While the authors are taking the lead in this effort, comments, additions, deletions, and questions from the broader statistical and user communities are essential to its ultimate acceptance and success as a Federal Standard. You can participate by accessing our *TOC* page on the WWW using the following URL:

http://www.census.gov/ftp/pub/std/TOC.html

11. References

[HTTP/1.0] Berners-Lee, T., *et al.*, "Hypertext Transfer Protocol — HTTP/1.0," 10/14/1995; URL <ftp://www.w3.org/hypertext/protocols/http/>.

[RFC-1738] Berners-Lee, T., *et al.*, "Uniform Resource Locators (URL)," 12/20/1994; RFC 1738.

[FGDC/SCDD-95] Federal Geographic Data Committee, Subcommittee on Cultural and Demographic Data (FGDC/SCDD), "Cultural and Demographic Data Metadata." Draft of May 1995.

[Kendall-60] Kendall, Maurice G. and William R. Buckland. *A Dictionary of Statistical Terms,*

[7] All dates are in 1996.

(2nd Ed.); 1960. Hafner Publishing Co.

[FIPS-123] National Institute of Standards and Technology. *Federal Information Processing Standard Publication 123: Data Descriptive File for Information Interchange (DDF)*. U.S. Department of Commerce; 1992. Adopts, with modifications, International Standard 8211-1985.

[FIPS-173] National Institute of Standards and Technology. *Federal Information Processing Standard Publication 173: Spatial Data Transfer Standard (SDTS)*. U.S. Department of Commerce; 1992.

[FIPS-192] National Institute of Standards and Technology. *Federal Information Processing Standard Publication 192: Application Profile for the Government Information Locator Service (GILS)*. U.S. Department of Commerce; 1994.

[OIRA-83] Office of Information and Regulatory Affairs. "Voluntary Standard: Procedures for Preparation of Abstracts for Public Use Statistical Machine-Readable Data Files" Office of Management and Budget; October, 1983.

[Rosen-91] Rosen, Bengt and Bo Sundgren. "Documentation for Reuse of Microdata from The Surveys Carried Out by Statistics Sweden"; Document 1991-06-28, Statistics Sweden.

[Sundgren-96] Sundgren, Bo *et al*., "Toward a Unified Data/Metadata Database at the Census Bureau," paper submitted to the Bureau of the Census Annual Research Conference, March 17-21, 1996, Arlington, Virginia.

[Szemraj-93] Szemraj, John A. and Billy R. Tolar. "Profile Development for the Spatial Data Transfer Standard"; paper delivered at *GIS/LIS*, November 2-4, 1993, Minneapolis, Minnesota.

[Yates-81] Yates, Frank. *Sampling Methods for Censuses and Surveys*, (4th Ed.); Charles Griffin & Co. Ltd, 1981, London & High Wycombe, England.

[a] **Acknowledgements**
The material in the proposed standard is based on the work of Bo Sundgren of Statistics Sweden [Rosen-91]. Principal contributors were Cheryl Landman of the CASIC Office; and Easley Hoy of the Statistical Research Division. Cathy Dippo of the Bureau of Labor Statistics provided extensive comments. The Cultural and Demographic Subcommittee of the Federal Geographic Data Committee gave us permission to use their work in a preliminary draft form.

[b] The SDTS currently consists of four parts:

Part 1 is a model of spatial phenomena, objects, and features. This part also contains the syntax and semantics for using the model. The rules for building any set of SDTS transfer files are also defined. These files consist of a metadata part and a data part.

Part 2 contains a list of attributive values and definitions. This is, the beginning of a thesaurus for metadata, i.e., descriptions of the data being transferred.

Part 3 provides a representation of an SDTS data file set, using the DDF.   This is a profile; a selection of particular aspects of many ISO and FIPS standards to be used in generating the transfer file.

Part 4, the *Topological Vector Profile (TVP)*, contains specifications for an SDTS profile for use with geographic vector data with planar graph topology.