TECHNICAL DEVELOPMENT OF THE PROPOSED STATISTICAL METADATA STANDARD

Gregory J. Lestina, Jr.,
William P. LaPlant, Jr.,
Daniel W. Gillman,
Martin V. Appel

Abstract

The Bureau of the Census is developing a Statistical Metadata Content Standard to define the necessary metadata to describe all aspects of survey design, processing, analysis, and data sets. The draft standards document must be easily reviewed by subject matter experts. Our experience has shown that information displayed in the format of a standard is not easily understood by experts outside the standards community. In order to facilitate review and discussion, we chose to display the standard in a format similar to a textbook's Table of Contents (TOC), with subsequent information displayed in an outline format. To further facilitate review and comment, the TOC was published in HTML format on the Bureau's World Wide Web server. Thus a reviewer, using internet browsers such as Mosaic or Netscape, can traverse the document, display context sensitive help, such as definitions, and leave behind comments as he/she scrutinizes the document.

In addition to displaying the standard, the TOC provides a mechanism for navigating and identifying subjects of interest about surveys for data dissemination, survey design and documentation, and process integration. The TOC also has certain system design implications. For example, it serves as a mapping between tools for data dissemination and integrated processing. In addition, the TOC is being used as a blueprint for building conceptual and logical data models for metadata repositories.

This paper will present a description of the TOC, a description of its uses, details of the Web implementation, and a demonstration of the system.

1 Introduction

The United States Bureau of the Census (BOC) is continually trying to improve timeliness and accuracy of the statistics provided to its customers. With the rapid advancement of the Internet and electronic processing of data, there are new and more efficient ways of managing and providing information to the Census Bureau's customers. In fact, these new advancements are further emphasized by the efforts of Vice President Al Gore. One effort, the National Performance Review (NPR), calls for government to improve service to the citizens, and a second effort is the development of the telecommunications National Information Infrastructure (NII), which includes widespread dissemination of government information over the national network. Therefore, our goal is to make data and documentation easier to access, understand, and use. The development and use of metadata standards and systems are essential to the success of this effort.

2 Statistical Metadata Definition

Statistical metadata is descriptive information or documentation about statistical data, i.e. microdata or macrodata. Statistical metadata facilitates sharing, querying, and understanding of statistical data over the lifetime of the data.

The types of statistical data (electronic or otherwise) are described as follows:

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.

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 procedures, such as sample designs, questionnaires, software, variable definitions, edit specifications, etc;

Administrative - the management information, such as budgets, costs, schedules, etc.

The systems, applications, and administrative components help to differentiate the sources and uses of statistical metadata (Gillman, Appel, LaPlant, 1996).

3 Current and Proposed Data Services Using or Providing Metadata

The need to review the current data dissemination systems began in the Fall of 1992 after a customer survey found that users of Census Bureau data are dissatisfied with the timeliness of data release, and the delivery of products and services as scheduled. The Reinvention Lab at the Bureau of the Census was organized to review these problems. The lab's initial focus was on post-data collection processing. After analysis of customer needs, members of the Lab came to the conclusion that they needed to frame the problem more broadly than just post-data collection processing and include all aspects of survey design and execution. The result was the development of the concept of the "Integrated Processing System" (IPS) (Reinvention Lab of the Census Bureau, 1994).

Since 1992, the Census Bureau has developed automated systems to improve the efficiency of statistical dissemination. The use of the World Wide Web for data dissemination has been a great success at the Census Bureau. An Internet prototype was begun in January, 1994, making the Census Bureau one of the pioneer federal agencies to disseminate data using the Internet. On average, the Bureau is now receiving more than 60,000 inquiries per day from customers who access this site. Through this Internet site, the BOC offers: population estimates and projections; economic indicators; international trade data; research from the Center for Economic Studies; news releases; state ranking and profiles, financial data for states, counties, cities, and school districts; and job vacancies. This prototype allows users to get Census information in seconds by using a software browser such as NCSA Mosaic or Netscape.

From this success, several programs are being developed using the Internet as a data access tool. The following four programs plan to or currently provide metadata and currently use or will use the Internet as a data access tool:

a) The Integrated Processing System (IPS) - The "Integrated Processing System" (IPS) is envisioned to be the "umbrella" for a compatible set of automated tools to design, conduct, and manage Census Bureau surveys and censuses in an effort to improve cost effectiveness, timely reporting, data quality, and data access (Reinvention Lab of the Census Bureau, 1994).

b) FERRET (Federal Electronic Research and Review Extraction Tool) - FERRET presently makes Current Population Survey (CPS) data and documentation available over the Internet through World Wide Web pages. A user is able to extract a dataset, review variable metadata, produce cross-tabulations, and display macrodata tables in SAS or ASCII format (Appel, Gillman, LaPlant, Creecy, 1996).

c) DADS (Data Access and Dissemination System) - The BOC plans to develop and implement an Internet based data access and dissemination system (DADS) initially focused on the 2000 Decennial Census and Continuous Measurement data sets, but with the ability to accommodate other data sets. The goal is to provide customers one general (electronic) system for all Census Bureau data access (Appel, Gillman, LaPlant, Creecy, 1996).

d) StEPS (Standard Editing and Processing System) - The objective of StEPS is to eliminate redundant processing by combining existing survey systems into one system. It is anticipated that 109 current surveys of the Economic Directorate will be migrating to StEPS by December 1999 (StEPS and the Economic Directorate View for Current Survey Processing, 1995).

4 Metadata Repository

The metadata repository under development will contain the metadata for survey design, processing, analysis, and data sets. Links to the data files themselves will eventually be made, creating a fully integrated data/metadata system.

Because the Bureau of the Census manages data in a decentralized and non-uniform way, the metadata repository will bridge the gap between the data and the users who wish to find them. The metadata repository will facilitate a solution for the data users while allowing the survey data managers to find a smooth transition to standard data management strategies.

There are many specific functions for which the metadata repository is being designed. Primarily, the metadata repository will be a standard tool for researchers and analysts to locate data and descriptions of surveys. Data dictionaries, record layouts, questionnaires, sample designs, and standard errors are examples of information that will be directly available.

Links from subject types, e.g., income, race, age, and geography, to data sets will allow users to locate data sets by subject. Less obvious, users can compare designs of different surveys and find common information collected by different surveys.

There are a number of types of users of data who require different kinds of metadata. Programmers will be interested in record layouts, data dictionaries, file storage medium, and other information needed to process data. Data analysts and researchers will have more interest in sample design, questionnaire, standard errors, and other similar information. Managers will be more interested in costs, schedules, and processes. A complete model of statistical metadata will have to take all these needs into account.

5 The Survey Design and Statistical Methodology Metadata Standard

This section describes the Survey Design and Statistical Methodology Standard (SDSM) and its importance in developing the metadata repository.

5.1 Elements and Structure of the Standard

The SDSM (Census Bureau, 1996, “Standard for Survey Design and Statistical Methodology Metadata”) was begun in January 1995 and is an extension of the Content Standard for Cultural and Demographic Data Metadata [FGDC/SCDD-95] of the Federal Geographic Data Committee (FGDC 1994a and FGDC 1994b). It is a standard intended to provide statistical metadata elements for administrative, planning, design, collection, analysis, and processing of statistical data and was developed primarily by Census Bureau researchers in consultation with subject matter experts and other researchers at the Census Bureau and Bureau of Labor Statistics. Textbooks and other standards were used as additional sources of information. The standard provides a way of classifying the metadata and classifying the meaning of the metadata. It is intended to provide a data user with the information to interpret and use statistical data.

The SDSM 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. It is designed to help the contributors and users of metadata answer questions such as “who”, “why”, “what”, “when”, “where”, and “how” for issues related to surveys, systems, and products. This outline or thesaurus is really a list of statistical metadata at the Census Bureau. From this list, we are able to define a logical business model for the Census Bureau. And from this logical business model, we are able to build the central metadata repository.

The metadata items of the SDSM standard are organized into "chapters” or “sections”. Each chapter or section represents a logical set of metadata. The inclusion rule (status of metadata items in the standard to ensure a complete set of documentation about the planning, collection, production and analysis of a data set) and definition are provided for each metadata data element. The following are the top-level chapters, their inclusion rules, and their definitions:

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. **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.

Each of the these chapters is subdivided into sections containing an outline of concepts. For example, Chapter 2, Planning is subdivided into Section 2.0 Point of Contact , Section 2.1 Project Conceptualization, Section 2.2 Design Proposal, and Section 2.3 Design Evaluation. Each of these sections may contain subsections. For example Section 2.1 contains Section 2.1.1 Options and Section 2.1.2 Shell or Model Design. There are approximately 533 chapters, sections, and sub-sections in the standard. These sections and subsections each contain a definition, citation, or maybe other forms of statistical metadata such as a URL. These metadata data elements also provide a basis for defining logical data models about how the Census Bureau does business. An example of an element in the standard is as follows:

2.1.1 **Options** *(optional).* <IPS>. Previous and related data and surveys that is considered in planning this project.

The SDSM standard does not specify the physical format of the content, the services to be provided, or the syntax to be used for its metadata data elements. For this reason, the SDSM is called a "*content* standard" because it defines which metadata items about statistical surveys and censuses are important. The SDSM supports labeling metadata content by tags included in the metadata itself or by indexes provided through tools (LaPlant, Lestina, Gillman, Appel, 1996).

5.2 Relationship to Other Standards

The CDDM is 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 (LaPlant, Lestina, Gillman, Appel, 1996).

6 The Table of Contents (TOC) for the SDSM Standard

In addition to the SDSM standard, a Table of Contents (TOC) to the standard is available. The TOC provides a map to the contents of the standard in a way that is similar to the table of contents of a book. The TOC provides a way for "readers" to quickly go to areas of interest (Census Bureau, 1996, "Table of Contents for Survey Design and Statistical Methodology Metadata").

The Table of Contents will be used with survey documentation tools 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. 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.

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)[1] 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.

The Table of Contents is an on-line summary of the Survey Design and Statistical Metadata Standard. The on-line TOC was developed to help reviewers of the SDSM. The on-line TOC presents the standard, in an easily accessible way, on the World Wide Web. The TOC is a combination of HTML pages and HTTP CGI[2] scripts

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

[2] 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. CGI scripts can provide a dynamic 2-way interface that is tailored to each user's unique request.

written in Perl[3]. 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 and allows users to enter their questions or comments on any of the elements in the standard.

The URL for the TOC is *http://www.census.gov/ftp/pub/std/www/TOC.html.* Through this site, we received many visitors, but received only a few comments to the definitions.

7 Applications

As mentioned earlier, the Table of Contents reflects the business processes of the Census Bureau and can be used to develop a metadata repository . Because of the popularity of new technologies that implement the World Wide Web, the Census Bureau must also re-examine the ways it collects, processes, and outputs its data and reports so that the process is more efficient for the data user. This section explains the Census Bureau's effort to model a metadata repository and explains the use of the TOC as an important tool in this model. When the model is complete, users will have the ability to view large holdings of Census metadata with considerable efficiency.

7.1 Modeling the Repository

The first step in this development involves defining the elements or concepts of the repository. For example, the Census Bureau is primarily concerned with concepts such as Design, Collection, Analysis, and Dissemination of statistics. These concepts are reflected in the Table of Contents.

The second step is to develop a logical model from the conceptual model (Barker, 1990). A logical model is a representation of how data is stored. A tool called Open Workgroup Repository (OWR)[4] is used to electronically create and implement the logical model, then build the repository. The repository uses a relational database for storing repository instances. We are using Oracle because of its availability and its ability to work with OWR. The OWR uses the Command Manipulation Language (CML) to translate logical schema to SQL. The OWR follows the standard provided by the Information Resource Dictionary System (IRDS), a standard that is used for implementing repositories.

7.2 Tool Development

As mentioned earlier, the Table of Contents is also a data dissemination and collection tool that is the center of metadata holdings at the Census Bureau. It is the central and connecting point for statistical systems located in many different physical locations at the Census Bureau. All documentation at the Census Bureau will be linked to the Table of Contents.

Section 4 discussed what the repository does for the user. By means of a CGI or web broker interface on the Internet, the Table of Contents is designed to link to the appropriate systems giving the user access to the documentation requested. All of these transactions are transparent to the user. The Table of Contents, therefore, provides access to all available documentation on the subject selected.

Not only is the Table of Contents useful in providing metadata to users, but it is useful for data providers in updating their data. An office or person that creates metadata, for example, may want to add the recent memorandum on a survey supplement to the repository. Or the same user may need to update a variable definition or add a new variable to the repository. This sort of interface would be provided interactively or as a batch process.

[3] An interpreted programming language.

[4] Produced by Manager Software Products, Lexington, Massachusetts

With the number of different Census Bureau programs, there could be numerous “front ends” or user interfaces with the Table of Contents, one for each statistical program. For example, a screen could be developed with keywords or for a user to input a word, and the document could be included with the appropriate section of the Table of Contents. The keyword search can also be used as a documentation search tool, with the Table of Contents providing reference to documents in the repository. Other front ends may include a menu system containing various subject areas of the Census Bureau referencing the Table of Contents. Or maybe the user prefers the idea of opening folders, as in Lotus Notes, to access Census Bureau files. These ideas will be developed at a later date.

7.3 Unifying Statistical Systems and Repositories

The Census Bureau is an organization containing many different programs that respond to many different customers. If data is processed differently for each program, it follows that the metadata for these programs are not coordinated. Creating a single metadata repository for the Census Bureau requires an effort to somehow coordinate the metadata for public and private access. The Table of Contents can be used to help coordinate these programs.

The statistical systems that collect, process, analyze, and disseminate statistical information, such as DADS, FERRET, IPS, and StEPS will be defined as Statistical Information Systems (SIS)(Gillman, Appel, LaPlant, 1996). These SIS allow users to access the data and metadata for data dissemination or automated survey processing. Our goal is to unify these SIS so that they are transparent to users looking for metadata from the central metadata repository. Unification of the systems will depend on the separate tables of contents of each of the other systems. These individual tables of contents would be used for mapping from one SIS to another and for mapping to the central repository (see Figure 1) (Gillman, Appel, LaPlant, 1996). These tables of contents would also provide an information outline for users and analysts.

There are a combination of various user interface tools that provide the user with the access to the data, metadata, and documentation. For example, a user can access and update metadata and a documentation library using SQL, SAS, word processing software, or various Internet tools.

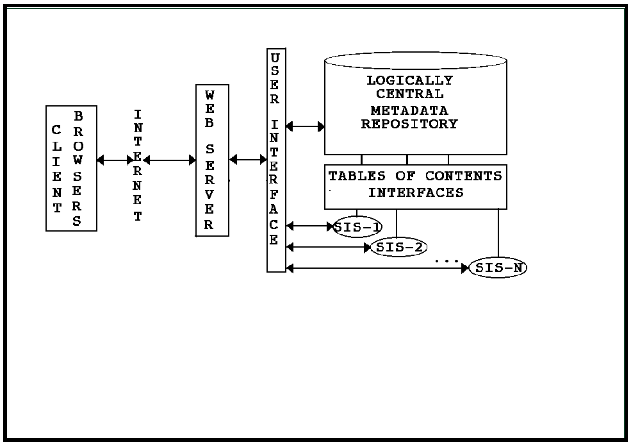
CLIENT
BROWSERS
INTERNET
WEB SERVER
USER INTERFACE
LOGICALLY CENTRAL METADATA REPOSITORY
TABLES OF CONTENTS INTERFACES
SIS-1
SIS-2
...
SIS-N

Figure 1. Unified System Architecture

As mentioned in section 5.1, the items in the individual tables of contents are “tags” and are used to associate items in other tables of contents. The tags are defined in the SDSM standard. Thus, requests for information can be easily transferred across systems.

8 Conclusion

Work on the logical data model for the metadata repository was begun in May, 1996. The Survey Design and Statistical Metadata Standard is due for final review in June, 1996. We are currently exploring ways to physically link items from the metadata repository to IPS, DADS, FERRET, and StEPS. We will then need to design tools to populate and manage the repository.

The task still facing us is to transform the logical interpretation of the Table of Contents to the physical access of the various SIS’. Unifying SIS’ through a logically central repository is expected to provide greater functionality than the sum of separate systems.

9 References

Appel, M. V., Gillman, D.W., LaPlant, W.P., Creecy, R.H. (1996), “Towards Unified Metadata Systems and Practices at the Census Bureau”, Integrated Statistical Information Systems (ISIS), May 1996, Bratislava, Slovakia.

Barker, Richard, Case*Method Entity Relationship Modelling, Addison-Wesley Publishing Company, Wokingham, England, 1990.

Census Bureau (1996), “Standard for Survey Design and Statistical Methodology Metadata” Draft Standard, Census Bureau internal document, in progress.

Census Bureau (1996), “Table of Contents for Survey Design and Statistical Methodology Metadata, Draft Standard”, Census Bureau internal document, in progress.

FGDC (1994a), Federal Geographic Data Committee, “Content Standards for Digital Gepspatial Metadata”, June 8, 1994.

FGDC (1994b), Federal Geographic Data Committee **-** Subcommittee on Cultural and Demographic Data, “Cultural and Demographic Data Metadata”, DRAFT, September 15, 1994.

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

[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.

Gillman, D.W., Appel, M.V., LaPlant, W.P. (1996), “Design Principles for a Unified Statistical Data/Metadata System”, 8th Scientific and Statistical Database Management Conference, June 1996, Stockholm, Sweden.

LaPlant, W.P., Lestina, G.J., Gillman, D.W., Appel, M.V. (1996), “Proposal for A Statistical Metadata Standard”, 1996 U.S. Bureau of the Census Annual Research Conference, March 1996, Arlington, Virginia.

Reinvention Lab of the Census Bureau (1994), “Integrated Processing System”, Systems Planning Document, December 15, 1994.

“StEPS and the Economic Directorate Vision for Current Survey Processing”, December 4, 1995