STATISTICAL METADATA MANAGEMENT: A STANDARDS BASED APPROACH

Daniel W. Gillman, Martin V. Appel, William P. LaPlant, Jr., Bureau of the Census[1]
Daniel W. Gillman, Statistical Research Division, Bureau of the Census

Key Words: Repository, Data Elements, IRDS

1. INTRODUCTION

Statistical metadata is the information and documentation needed to describe and use statistical data sets for the lifetime of the data. Too often, this information is scattered, incomplete, or missing. Many times the only source for some information is from subject matter experts.

The effective and efficient management of statistical metadata greatly increases the usefulness of statistical data. Since metadata is data, it can be stored and retrieved in a repository just as the data it describes is stored and retrieved in a database. Statistical metadata can also be used to facilitate survey design, processing, management, and analysis. It is the electronic storage and organization of statistical metadata which will allow statistical agencies to develop automated survey design and processing systems. So, the organized use of metadata enables statistical agencies to conduct their programs in ways that were not possible before.

The Bureau of the Census (BOC) is conducting research into the collection, content, storage, and delivery of statistical metadata. The main focus of the research is the design and implementation of a prototype logically central statistical metadata repository for use with Internet data dissemination systems and automated integrated survey processing systems. Integrating the components of metadata management requires careful planning. Standards are the vehicle which will enable this work. International, American, U.S. Government, and BOC standards are all being brought to bear to solve these problems.

This paper will describe the purposes of a statistical metadata repository, an architecture for developing a prototype at the BOC, the standards which will be used to guide the development of the prototype, and how the architecture and standards address the problems of metadata management.

2. DEFINITIONS

Statistical Metadata[2] is descriptive information or documentation about statistical data, i.e. microdata and macrodata. Statistical metadata facilitates sharing, querying, and understanding of statistical data over the lifetime of the data.

The two types of statistical data (electronic or otherwise) are described as follows (see Lenz, 1994):

- **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 (see Sumpter, 1994) 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 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.

Some authors (see, for example, Sundgren, 1991b, 1992, 1993) refer to the applications and administrative components of metadata as meta-information. We chose to use the term metadata because it seems to simplify the discussion.

Statistical metadata and metadata repositories have two basic purposes (see Sundgren, 1991a, 1991b, 1992, 1993):

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

A potential end-user of statistical information needs to

- identify,
- locate,
- retrieve,
- process,
- interpret, and
- analyze

statistical data that may be relevant for a task that the user has at hand.

The production-oriented user's tasks belong to the following types of activities:

- planning/design/maintenance,
- implementation/processing/operation, and
- evaluation.

An **input-oriented** statistical agency is one where the statistical surveys they conduct or manage are also the natural building blocks of its organization. The BOC is

currently an example of such a statistical office.

An **output-oriented** statistical agency is one which focuses on meeting the needs of its customers. The BOC is striving to become more output-oriented. See Sundgren (1991a, 1991b, 1992, 1993) for a more detailed discussion of these ideas.

Output-oriented database systems relate data from different surveys. They need special software and metadata tools for reconciling data from different sources and for helping the users to interpret and analyze the data. This paper describes the pieces necessary to build those metadata tools.

Statistical Metadata Repository (MDR) is the project to build a logically central statistical metadata repository. Its design will be based on three main standards, which will be discussed later. The MDR is intended to be a source of metadata for all the BOC programs so comparisons of designs, processing, analysis, or data can be made across time and survey programs. The MDR will support both Internet data dissemination tools and automated integrated survey processing systems. A proof-of-concept system has been built (see Gillman and Appel, 1994)..

Statistical Design and Survey Methodology Metadata Content Standard (SDSM) is a standard under development at the BOC to specify the metadata necessary to describe survey designs, processing, analyses, and data sets completely. Approval for the standard will be sought through the formal standard development procedures of the BOC, and the process is expected to be completed by September 1996. This standard is one of the three which form the basis for the design of the MDR, and will be discussed in more detail in section 5.

3. CURRENT METADATA TOOLS

The BOC does have some metadata tools and metadata-driven systems. Metadata-driven systems are electronic data dissemination or automated integrated survey processing systems which use metadata to guide the user, providing choices and information at each step. Most of the systems are primarily end-user oriented, and some (StEPS, IPS, etc.) are primarily production oriented. The statistical metadata repository will support both purposes. All the systems which are in production are end-user oriented.

Some of the important metadata tools and metadata-driven systems that exist at the BOC or are under construction are briefly described below.

Automated Reference Rack (ARRk) is a hypertext based system designed using Lotus SmartText. Telephone clerks at the BOC use ARRk to help the public find appropriate published data by searching for key words across short descriptions of each available file. Descriptions contain information about subjects, coverage, storage medium, cost, and ordering information.

BOX Files is a mechanism for incorporating metadata into ASCII data files (see Bean, 1991). The major advantage of this system is that the systems level metadata that describes each file are automatically carried along with the data in a file. The BOX file format is the basis for a recently issued BOC information technology standard for archiving data, although this decision is under review.

Data Extraction System is a general metadata-driven system for extracting data from master data sets into one of several popular data formats, including SAS data sets. It also uses files stored in the BOX format (see above).

Surveys-On-Call is an on-line system for accessing publicly available Survey of Income and Program Participation (SIPP) and Current Population Survey (CPS) data. This system is currently in production and is accessible over the Internet (via the BOC home page) and by modem. Surveys-On-Call is a UNIX based menu system which is an implementation of the Data Extraction System.

Extract (see Zeisset, 1993) is a metadata-driven system in use with CD-ROM products sold by the BOC. Libraries and other public facilities often have these products on the shelf. Extract is a menu driven dBase application which allows the user to construct and extract data files. Metadata is also available and can be appended where necessary in the extracted files.

CENSAS is a project for an automated data and information delivery system based on Decennial Census data for both internal and external customers. Extracts are delivered to the user as SAS datasets to which the user can apply the desired tallies or other statistical analyses. A Beta Test version is available.

CENDATA is an on-line BOC data system containing current and historical data, both demographic and economic. Examples include Foreign Trade Statistics, Quarterly Financial Reports, County Business Patterns, Wholesale and Retail Trade, Center for International Research, Agriculture County, and Manufacturers, Shipments, Inventories, and Orders Survey data.

FERRET (Federal Electronic Research and Review Extraction Tool) (see Capps, 1995) is a metadata-driven data extraction tool available on the Internet that allows users to find information about monthly demographic survey data using a World Wide Web browser. Users can select microdata or macrodata and download files as SAS data sets or ASCII files.

StEPS (Standard Economic Processing System) (see StEPS, 1996) is an integrated survey processing system the objective of which is to eliminate redundant processing by combining existing survey systems into one system. The scope of the StEPS system includes access to basic survey processing functions and some additional functions.

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

framework for the integration of generalized processing system components with data collection and other tools.

DADS (Data Access and Dissemination System) is the name for the BOC initiative to develop and implement data access and dissemination focused on the 2000 Decennial Census and Continuous Measurement data sets, but with the ability to accommodate other data sets having geographic detail, such as those produced from the Economic and Agricultural Censuses. The main objective of DADS is to provide one general (electronic) system for all access to BOC data.

4. STATISTICAL METADATA REPOSITORY

The MDR is being designed to assist with two new types of tools which are under development at the BOC: Internet data dissemination (e.g. DADS, FERRET); and automated integrated survey processing systems (e.g. IPS, StEPS). These tools correspond to the end-user oriented purpose and production oriented purpose, respectively, of statistical systems. Statistical systems are known formally as **Statistical Information Systems (SIS)** (see Sundgren, 1991b, 1992, 1993; or Gillman, Appel, and LaPlant, 1996).

The current goal of the MDR project is to build a prototype for a standards based logically central metadata repository. The **Open Workgroup Repository (OWR)** software (v3.0) has been purchased from **Manager Software Products (MSP)**. The OWR product is based on the **Information Resource Dictionary System[3] (IRDS)** standard (see NIST, 1989). IRDS is the second standard which forms the basis for the design of the MDR, and will be discussed in more detail in section 5.

4.1 Purposes For A Statistical Metadata Repository

The eventual plan for the MDR is that it will contain the metadata for survey designs, processing, analyses, and data sets for all surveys the BOC performs. Links to the data files, documentation, and images (such as questionnaire forms) will be made (see Sundgren, *et al,* 1996; or Appel, *et al,* 1996).

Because the BOC manages data in a decentralized and non-uniform way, the MDR will bridge the gap between the data and the users who wish to find them. On one hand there is a need for the managers for each survey to create and manage their data in the most efficient way for their processing needs. On the other, there is a need for data users to be able to find and access data efficiently and effectively. The MDR 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 functions for which the MDR is being designed. Primarily, the MDR 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 obviously, users can compare designs of different surveys and find common information collected by different surveys.

The MDR will help facilitate data administration at the BOC. Many surveys define data elements with the same name but with (slightly) different definitions. An aim of the MDR is to help people manage this problem. If definitions and other attributes of data elements are standardized across surveys, through the use of a data element registry (a subset of MDR), then confusion generated by the differences in meaning will be reduced. Naming standards and conventions are also needed to reduce the confusion. The MDR will provide the information necessary for the user to understand the distinctions and similarities among data elements from multiple data sources. The design of the data element registry part of the MDR will be based on a standard, the third, and last, standard which will form the basis of the design of the MDR. It, too, will be discussed in more detail in section 5.

Many of the purposes for the MDR are associated with both the end-user orientation and production orientation. Here we will list the end-user oriented purposes. The typical end-user oriented SIS is an Internet data dissemination system. Some of the major functionality for the MDR in support of this is:

- Location of data sets by survey name and date or content (e.g. household income);
- Names, definitions, and related information about data elements and links to the surveys and data sets that use them;
- Links to documentation describing aspects of survey design, processing, or analysis;
- Links across documents to identify common themes contained in them;
- Links to images (e.g. questionnaire forms) that are of interest;
- The ability to search the information potential through query languages such as **SQL**.

The typical production oriented SIS is an automated integrated survey processing system. Most of the purposes of the MDR for the end-user oriented systems are common to the production oriented systems as well. Often, production oriented SIS users will be survey analysts working within the BOC (statistical agency). They have and need access to confidential data which external end-users cannot have access to. The additional functionality must support this use, such as:

- Links to all the data sets produced by the instance of a survey (e.g. Current Population Survey, June 1996);
- Links to frame, sample, and administrative records files;
- Links to a management information system;
- Links to some confidential metadata such as disclosure analysis algorithms.

These lists are not meant to be inclusive, but to give a fairly extensive picture of the potential uses for the MDR.

4.2 Architecture For A Statistical Metadata Repository

As mentioned above, the MDR will be built using MSP's OWR software. The software provides a developer platform which employs an underlying **relational database management system (RDBMS)**. Oracle version 7 is the RDBMS which is in use. The data (in this case metadata) stored in the repository is managed by the RDBMS. The developer platform provides an **entity-relationship-attribute (ERA)** modeling capability. Every OWR repository is defined by a model, and the software provides the tools for developing and testing the model. Another part of the software translates the model and any data to an application in the RDBMS. A special language defined in the IRDS standard called **Command Manipulation Language (CML)** is responsible for this translation. CML can be used to define and update the model; and add, change, or delete data. This paradigm is a very flexible one for developing the MDR.

The MDR modeling effort is currently underway. There will be several dimensions to the model, each one considered as a separate model in itself. They are listed here.

- **Business data model** - A prototype of this model has been built[4]. It defines the metadata and relationships necessary to describe the business (surveys and survey data) of the BOC.
- **Data Element Registry** - A prototype of this model is being built. It is based on work of the **ANSI Standards Committee X3L8 - Data Representations**. A data element registry is a mechanism for managing the names, definitions, and other attributes of data elements. An implementation should provide easy access and links to the information about an organization's data elements. See section 5 for more details.
- **Metamodel** - This is the model describing schemas of information about data sets, such as specifications for record layouts or database schemas. It also contains versioning, security, user type, and search category information. The metamodel is the repository's view of the business data model and the data element registry.

The integrated model will be used to build a repository in OWR. We will refer to this integrated model as the **integrated statistical metadata (ISM)** model.

Repository access and update tools will have to be built to let other tools, analysts, designers, and researchers use the repository. The repository models will help determine how these tools are designed. This will be discussed further in section 6.

5. APPLICABLE STANDARDS

In this section the applicable standards which will be used to guide the development of the MDR and its associated tools will be described. Each of these standards have already been mentioned previously. Some additional standards which will be appropriate to the design of the MDR will also be briefly described.

5.1 Information Resource Dictionary System (IRDS)

IRDS is a standard which addresses the use, control, and documentation of the information resources of an enterprise (see NIST, 1989). It is an application of another standard, **Reference Model for Data Management (RMDM)** (see ISO, 1995). RMDM specifies a series of interlocking pairs of databases and schemas, where (informally) the schema stored in the database at one level defines the database stored at the next lower level. See figure 1.

IRDS uses three interlocking level-pairs of the RMDM specification. The levels are as follows:

- **application level** - which is the survey data itself;
- **dictionary level** - which describes the actual data in an application, i.e. contains the schemas for the databases containing survey data;
- **dictionary definition level** - which describes the schemas used at the dictionary level, i.e. contains a description of how the schemas fit together (the MDR ISM model);
- **definition schema level** - which describes the schemas used at the dictionary definition level, i.e. contains the modeling paradigm for the MDR ISM model (the meta-metamodel).

The dictionary level is all the record layouts and database schemas in use at the BOC for describing its survey data sets at the application level. The ISM model (see section 4) is a scheme for linking the various record layouts and database schemas together. Finally, there is a level (the definition schema level) for describing rules for building the ISM model. The OWR software tool itself has much of this meta-metadata as part of its overall design. For instance, the E-R-A modeling paradigm and the CML interface language are constituents of the meta-meta level.

The level-pairs are each of the three pairs of levels going down the list above:

- application - dictionary
- dictionary - dictionary definition
- dictionary definition - definition schema.

These pairs can be viewed in a contextual frame (see Graves and Gillman, 1996) by observing that the application - dictionary pair is the "operational" context, i.e. the operations of the BOC. The dictionary - dictionary definition pair is the "design and maintenance" context, i.e. the business activities of the BOC (c.f. the business data model and the data element registry). Finally, the dictionary definition - definition schema pair is the "policy and standards" context, i.e. standards, methods, and rules the BOC uses to determine the way it conducts business. See figure 2[5].

5.2 Data Element Standards

Data elements (or variables) are the fundamental units of data an organization collects, processes, or

disseminates. A data element registry is a mechanism for managing data elements in a logical fashion. Data element registries organize information about data elements, provide access to the information, facilitate standardization, help identify duplicates, and facilitate data sharing.

Data element registries are like data dictionaries in that they contain definitions of data elements. But more than data dictionaries, they contain all the information about individual data elements that an organization requires. Data dictionaries are usually associated with single data sets (files or databases), but a data element registry contains information about the data elements for an entire program or organization.

The information contained in a data element registry is part of an organization's metadata. Therefore, the registry itself will be part of MDR (see section 4). When the information about data elements is organized into a registry or repository, it can be made available to a variety of people and processes.

Important applications for data element registries include SIS's. Electronic data dissemination requires easy access to information about data elements. Data element names, definitions, and classification schemes will help users in locating and understanding data sets.

To design new automated integrated survey processing systems that will include sample and questionnaire design, automated edits and imputation, and coding systems, full descriptions of data elements are required. Designers need to know the definitions of all variables that may be affected by the programs they are using.

The model for the data element registry will be based on the conceptual framework contained in the draft ANSI standard, **The Metamodel for the Management of Shareable Data (MMSD)**, ANSI X3.1125-D. It, in turn, incorporates all the principles described in an emerging international standard, **Specification and Standardization of Data Elements**, ISO/IEC 11179 (see ANSI X3L8, 1996). It provides a conceptual model for building a data element registry and it contains some additional functionality which is beyond the scope of the underlying international standard. A complete data dictionary describing all the entities, attributes, and relationships of the conceptual metamodel is included in this document. See the Appendix for further details about ISO/IEC 11179.

The MMSD metamodel provides a detailed description of the types of information which should belong to a data element registry. It provides a framework for how data elements are formed and the relationships among the parts. Implementing this scheme will provide users the information they need to understand an organization's data elements.

5.3 Survey Design and Statistical Methodology Metadata Content Standard

The **Survey Design and Statistical Methodology Metadata Content Standard (SDSM)** (see LaPlant, *et al*, 1996; or Census Bureau, 1996a) is a proposed statistical metadata content standard for the BOC. It will provide a description of the information or documentation about statistical data. It is intended to be a comprehensive 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. The SDSM provides a mechanism for comparing and linking among statistical models (e.g. ISM model), obtaining consensus on statistical concepts independent of how those concepts are named. SDSM will provide developers and users of statistical products with a common vocabulary for describing the design processing, analysis, and data sets for censuses and surveys. The SDSM also will serve as a glossary of statistical metadata concepts. 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.

The other main purpose of the SDSM is to provide the content (entities and attributes) for the development of the business data model. It is organized into chapters which correspond to the main entities of a business process model previously developed at the BOC (see Reinvention Lab, 1994).

Each section in the SDSM 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 electronically. This last description must be provided as a **Uniform Resource Locator (URL)** (see LaPlant, *et al.*, 1996)

A table of contents (TOC) outline (see Census Bureau, 1996b) of the SDSM has been developed. It was patterned after work done at Statistics Sweden (see Rosen and Sundgren, 1991) and originally built to provide a "user-friendly" view of the standard. There are two other uses that are being developed for it, too: 1) to be used as a "check list" for users who need to provide metadata or users who want to search metadata from the MDR; and 2) to serve as an interface between the MDR and other tools which need to share metadata (see Gillman, Appel, and LaPlant, 1996). These will be discussed in more detail in section 6.

5.4 Other Standards

There are a number of other standards which will be of help in the design and use of the MDR. These include data interchange standards such as the **Spatial Data Transfer Standard (SDTS)**, FIPS-173, (see FGDC, 1994); the US indexing system service standard, **ANSI/NISO Z39.50**; and its profiles for automated indexing on the Internet, the **Government Information Locator Service (GILS)**, FIPS-192, and proposed and existing **Wide Area Information Services (WAIS)**.

The SDSM is intended to be used with the

Standard for Cultural and Demographic Data Metadata (CDDM) draft (see FGDC, 1995). The CDDM, is in turn, mapped to the metadata portions of the SDTS and supports providing metadata for the GILS. The SDSM assumes the existence of these other standards which define additional, related, metadata. The thematic content of a data file is provided by the CDDM while the physical layout is provided by either the SDTS and the **Data Descriptive File for Information Interchange (DDF)**, FIPS-123, specification or by GILS.

The SDTS is designed to assist in moving the contents of Geographic Information Systems (GIS) databases between GIS servers or to exchange data between systems that have the capability of generating, analyzing, storing, or displaying geographic data, such as a SIS. This is important because most of the data collected by the BOC has a geographic component to it.

GILS is implemented through a decentralized collection of servers and formatted information records contained on those servers. The Z39.50 protocol is used to make the information available on the Internet. These services will be used by the public to find information throughout the federal government by using Internet search and retrieval tools.

These standards will all be brought to bear in the development of the ISM model for the MDR.

6. METADATA MANAGEMENT

The main aspects of managing metadata are content, storage, collection, and delivery. This section will describe how the standards based approach and the proposed design architecture address each of these aspects.

6.1 Content and Storage

Content refers to the metadata which will be collected and stored in the MDR, and **Storage** refers to how, i.e. the physical and logical mechanisms for storing the metadata. Much of the paper to this point has been addressing these issues.

The prototype MDR is being built using the OWR software from MSP. OWR uses Oracle RDBMS as its underlying storage mechanism and is based on the IRDS standard. The interlocking level-pairs of IRDS present a framework for organizing the statistical metadata of the MDR. The ISM model is a logical view of how the statistical metadata will be organized in the MDR, and it fits into the level-pair framework of IRDS. The SDSM, MMSD, and other standards (e.g. CDDM) define the content which is used to design the ISM. Also, the metamodel currently under development is a higher level view of the other parts of the model.

6.2 Collection and Delivery

The **Collection** and **Delivery** refer to creating, replacing, updating, deleting, and querying metadata in the MDR. Delivery (querying) will be part of the design of SIS's which work with the MDR. User interfaces for metadata-driven systems will let users query the metadata to locate data or other survey information. Query languages such as SQL (for RDBMS's) and CML (for OWR) will allow the user to retrieve any metadata which is in the MDR. Other search mechanisms such as WAIS, key word, and hyper-text are available through the Internet. This is especially important for documentation databases. MSP is releasing an **API (application program interface)** for OWR which will allow automated queries from user interfaces such as Internet browsers.

The TOC view of the SDSM (see section 5) can be used as a check list for categories of metadata. For users who are wishing to find information about surveys, searching the TOC for the appropriate subject (e.g. questionnaire design) will be useful. Since the SDSM is designed to be a complete description of survey design, processing, analysis, and data sets, then the TOC view will provide users access to all the metadata the BOC has about a survey.

From the metadata collection perspective, the TOC provides the survey designer or analyst a checklist and a place holder for metadata that they need to provide. Metadata collection tools will automatically update the MDR with information provided by the designer/analyst; the attributes in the ISM model (the MDR) are determined by the section of the TOC which has been selected.

Several prototype metadata collection tools are in place at the BOC and other statistical agencies. SCBDOK (at Statistics Sweden), Document Management System (DMS - in use with FERRET at BOC), and the commercial document management system PCDOC (for 1997 Economic Censuses) are all designed or being designed under the framework outlined above.

Metadata collection is recognized as a very difficult problem because of the fundamental changes that the survey design and analyst teams must go through to perform their work. At the BOC and other statistical agencies, metadata (mostly documents, often in the form of memos) is created either electronically or on paper for each survey, but it is just beginning to be stored in an organized repository, database, or document management system. Asking people to use a new system to capture this metadata and organize it represents a big change. The tools that are created must mimic as closely as possible the working paradigm already in place, such as the use of certain word processors and templates for creating documents. A major problem is that the working paradigm for each survey design and analysis team is different. So, creating common tools will require substantial planning. Also, incentives must be found so that the designer/analysts will want to provide the metadata to the MDR. No matter how well designed, tools without an obvious payoff to the user will not be used. Management can help with the adoption of metadata collection tools by supporting their use, but the end-users will ultimately decide their fate.

6.3 System Integration

In addition to the tools for collecting and delivering metadata, the integration of the MDR with other SIS's needs to be seamless. Again, the TOC can be used.

Each SIS has a metadata component which is based on a model of the information. Usually these models are not as extensive as the ISM model for the MDR, but the MDR must be able to communicate with these SIS's in a seamless way. The TOC can act as a map between the MDR and the other systems. There is a map from the MDR to the ISM model (the ISM was designed that way), and maps can be built from the TOC to the other systems by mapping the TOC their metadata models. Then, a map will exist from the MDR to each SIS, through the TOC. The MDR will act as a hub, a central communication link between the different SIS's in use at the BOC (see Gillman, Appel, and LaPlant, 1996).

6.4 Data Administration

The adoption of data element standards by the BOC for use with the MDR will require more than supplying information about data elements. Data administration is the active management of the information about all the agency's variables. No function of this type exists at the BOC at this time at the agency level.

The data element standards described above (see section 5.2) detail the information that is required for accurate and complete data administration. They also describe the organization of a data element registry for that information. Of course, the registry will be part of the MDR.

There is a human side to data administration which is lost in the discussion of the MDR, however. Some of these functions are listed here:

- Designing and implementing naming conventions;
- Designing rules for forming data element definitions;
- Determining which data elements have the same meanings as others;
- Ensuring all necessary information is properly supplied for each registered data element;
- Working with data administrator of other agencies to facilitate the sharing of data.

Data administration will require a large commitment from the BOC, but it will greatly enhance the usefulness of BOC data, make the MDR a better tool, and facilitate the sharing of data between groups within the BOC or with other agencies.

7. CONCLUSION

This paper has discussed the BOC architecture for building a statistical metadata repository (MDR) using standards developed by international, national, and U. S. Government organizations. Included is the BOC itself. The three most important standards which support this development (IRDS, MMSD, and SDSM) were described in detail, and their relationship to the current modeling efforts were described.

The MDR will not be an end in itself. Instead, it will work in conjunction with Internet data dissemination and automated integrated survey processing tools. Several examples of both of these tools are under development at the BOC. The MDR prototypes must be ready in time to meet the schedules of these other tools.

Work to build MDR prototypes has begun. In addition to the modeling work, access and update tools are under development, too. The MDR must be made to work seamlessly with the tools it supports and as a bridge between disparate tools which use it as a metadata source.

A schedule for development of a prototype for the MDR has not been set, but one must be ready in time for the DADS prototype of the 2000 Census Dress Rehearsal in 1998.

8. REFERENCES

• Appel, M. V., Gillman, D. W., LaPlant, W. P. Jr., Creecy, R. H. (1996), "Towards Unified Metadata Systems and Practices", ISIS-96, Bratislava, Slovakia, May 21-24, 1996.

• ANSI X3L8 - Data Representations (1996), "ISO/IEC 11179 Part 1 - Framework for the Specification and Standardization of Data Elements, Working Draft 7", February 1996.

• Bean, E. C. Jr. (1991), "ASCII BOX Files for Data Portability", Census Bureau internal document, Office of the Director - Demographic Programs.

• Capps, C. (1995), "Overview of the Technical Architecture for FERRET", Census Bureau internal document, Demographic Surveys Division.

• Census Bureau (1996a), "Statistical Design and Survey Methodology Metadata Content Standard", Draft, Census Bureau Internal Document, July 2, 1996.

• Census Bureau (1996b), "Table of Contents for Statistical Design and Survey Methodology Metadata Content Standard", Draft, Census Bureau Internal Document, July 2, 1996.

• FGDC (1994), Federal Geographic Data Committee, "Content Standards for Digital Geospatial Metadata", June 8, 1994.

• FGDC (1995), Federal Geographic Data Committee - Subcommittee on Cultural and Demographic Data, "Cultural and Demographic Data Metadata", Final Draft, January 18, 1995.

• Gillman, D. W. and Appel, M. V. (1994), "Metadata Database Development at the Census Bureau", Presented at the UN/ECE METIS Working Group Meeting, Geneva Switzerland, November 22-25, 1994.

• Gillman, D. W., Appel, M. V., and LaPlant, W. P. Jr.

(1996), "Design Principles for a Unified Statistical Data/Metadata System", Proceedings of SSDBM-8, Stockholm, Sweden, June 18-20, 1996.

• Graves, R. B. and Gillman, D. W. (1996), "Standards for Management of Statistical Metadata: A Framework for Collaboration", ISIS-96, Bratislava, Slovakia, May 21-24, 1996.

• ISO (1995), "Reference Model for Data Management", ISO/IEC 10032:1995(E).

• LaPlant, W. P. Jr., Lestina, G. J. Jr., Gillman, D. W., and Appel, M. V. (1996), "Proposal for a Statistical Metadata Standard", Census Annual Research Conference, Arlington, VA., March 18-21, 1996.

• Lenz, H.-J. (1994), "The Conceptual Schema and External Schemata of Metadatabases", Proceedings of SSDBM-7, pp160-165, Charlottesville, VA, September 28-30, 1994.

• NIST (1989), National Institute for Standards and Technology, "Information Resource Dictionary System (IRDS)", Federal Information Processing Standard (FIPS) Publication 156, April 5, 1989.

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

• Rosen, B. and Sundgren, B. (1991), "Documentation for Reuse of Microdata from the Surveys Carried Out by Statistics Sweden", Research and Development Statistics Sweden, June 28, 1991.

• StEPS (1996), "Standard Economic Processing System Document 1: Concepts and Overview", Draft 2, Internal Census Bureau Document, April 16, 1996.

• Sumpter, R. M. (1994), "White Paper on Data Management", Lawrence Livermore National Laboratory document, 1994.

• Sundgren, B. (1991a), "Towards a Unified Data and Metadata System at the Australian Bureau of Statistics - Final Report, December 2, 1991.

• Sundgren, B. (1991b), "Statistical Metainformation and Metainformation Systems", R&D Report Statistics Sweden, 1991:11.

• Sundgren, B. (1992), "Organizing the Metainformation Systems of a Statistical Office", R&D Report Statistics Sweden, 1992:10.

• Sundgren, B. (1993), "Guidelines on the Design and Implementation of Statistical Metainformation Systems", R&D Report Statistics Sweden, 1993:4.

• Sundgren, B., Gillman, D. W., Appel, M. V., and LaPlant, W. P. (1996), "Towards a Unified Data and Metadata System at the Census Bureau", Census Annual Research Conference, Arlington, VA., March 18-21, 1996.

• Zeisset, P. T. (1993), "Meta-Information for Summary Statistics: The EXTRACT Experience", Proceedings of Statistical Metainformation Systems Workshop, Luxembourg, February 2-4, 1993.

9. APPENDIX

ISO/IEC 11179 is an emerging international standard being built in six parts. Parts 3, 4, 5, 6 are all accepted international standards. Parts 1 and 2 are scheduled to be standards in the summer of 1997. A short description of each part follows here.

The Framework for the Specification of Data Elements, Part 1 of ISO/IEC 11179, introduces and discusses fundamental ideas of data elements essential to the understanding of the set of standards and provides the context for associating the individual parts of the International Standard.

Classification for Data Elements, Part 2 of ISO/IEC 11179, provides procedures and techniques for associating data element concepts and data elements with classification schemes for object classes, properties, and representations - the constituent parts of a data element. These procedures and techniques shall assist Registration Authorities in applying classification schemes which enable them to perform activities such as:

- analyze object classes, data element concepts, and data elements;
- make comparisons within the following categories: object classes, data element concepts, and data elements;
- reduce the variety of data element concepts and data elements;
- define and identify data element concepts and elements unambiguously;
- assist in the analysis of data elements for the purpose of assigning registration status;
- retrieve data element concepts and data elements from a data register;
- recognizing relationships among data element concepts and data elements.

Basic Attributes of Data Elements, Part 3 of ISO/IEC 11179, specifies attributes of data elements. It is limited to a set of data element basic attributes, independent of their usage in application systems, databases, data interchange messages, etc. The basic attributes specified are applicable for the following main activities:

- definition and specification of the contents of data

element dictionaries;
- design and specification of application-oriented data models, databases, and messages for data interchange;
- actual use of data in communications and information processing systems;
- interchanging or referencing among various collections of data elements.

Rules and Guidelines for the Formulation of Data Definitions, Part 4 of ISO/IEC 11179, provides guidance on how to develop good data element definitions. A number of specific rules and guidelines are presented in this document that specify exactly how a data element definition should be formed. A precise, well-formed definition is one of the most critical requirements for shared understanding of a data element; well-formed definitions are imperative for the exchange of information. Only if every user has a common and exact understanding of the data element can it be exchanged trouble-free.

Naming and Identification Principles for Data Elements, Part 5 of ISO/IEC 11179, provides guidance for the identification of data elements. Identification is a broad term for designating, or identifying, a particular data element. Identification can be accomplished in various ways, depending upon the use of the identifier. Identification includes the assignment of numerical identifiers, or registration identifiers, that have no inherent meanings to humans; icons (graphic symbols to which meaning has been assigned); and names with embedded meaning, usually for human understanding, that are associated with the data element's definition and value domain.

Registration of Data Elements, Part 6 of ISO/IEC 11179, provides instruction on how a registration applicant may register a data element with a central Registration Authority and the allocation of unique identifiers for each data element. Maintenance of data elements already registered is also specified in this document.

10. ENDNOTES

1 The views reflected in this paper are attributable to the authors and do not necessarily represent those of the Census Bureau.

2 Defined with assistance from participants at Statistical Metadata Workshop, November 14-15, 1995, Bureau of Labor Statistics, Washington, DC.

3 ANSI X3.138 and FIPS 156

4 Database Design Solutions of Bernardsville, NJ was hired to help with this modeling effort.

5 Taken from Graves and Gillman, 1996.

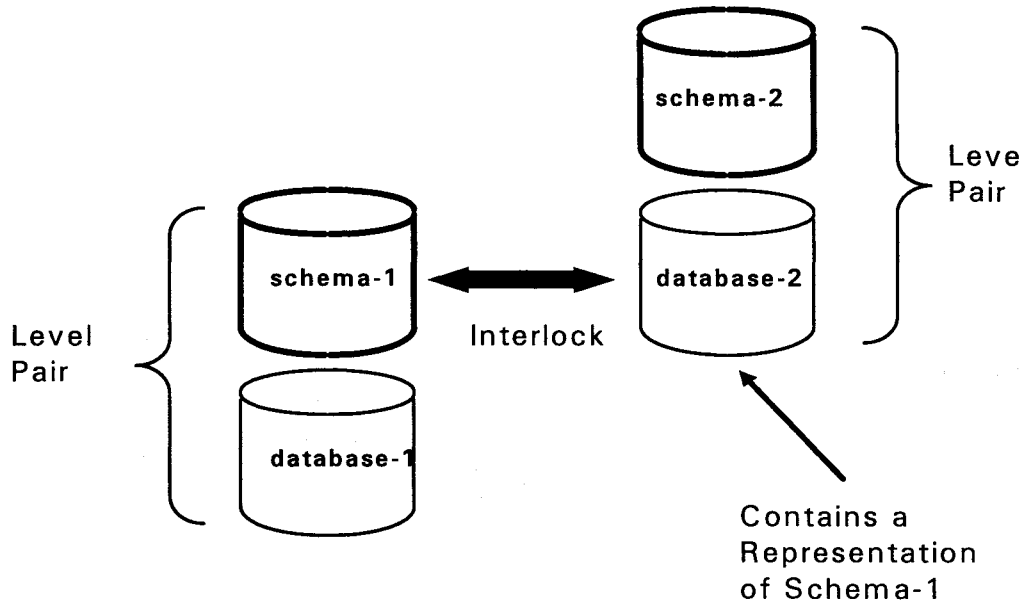

Figure 1: Interlocking Pairs of Reference Model for Data Management

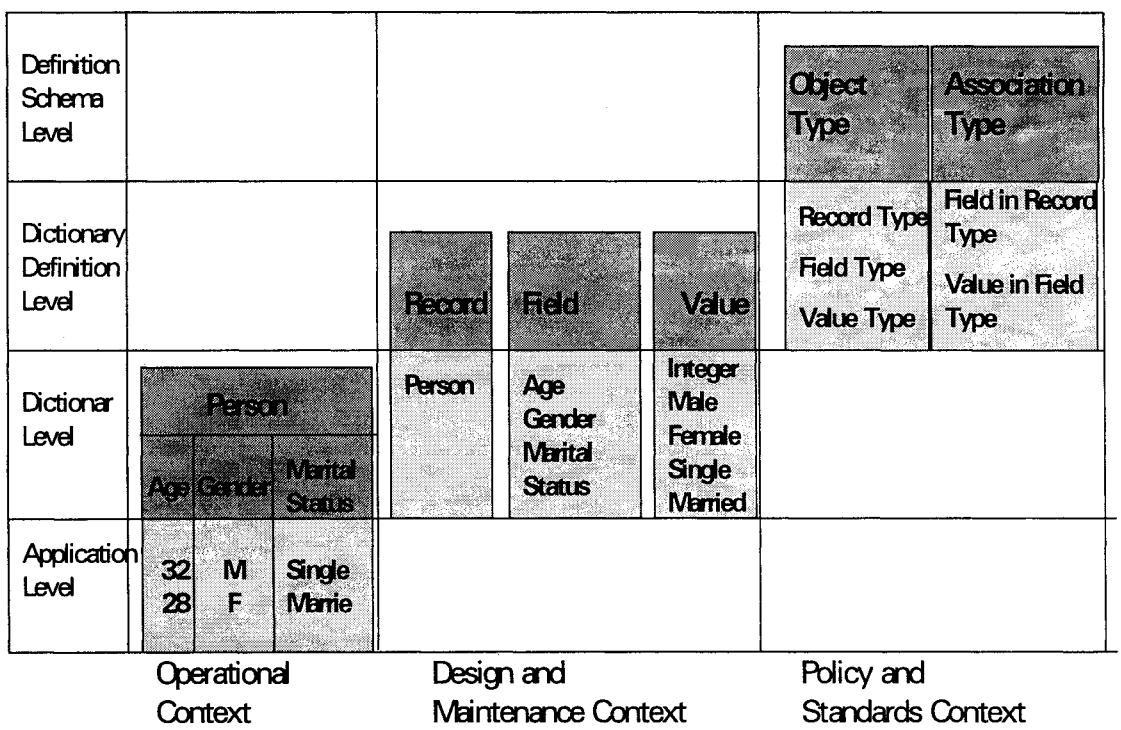

Figure 2: Statistical Example of IRDS Framework