Statistical Metadata Research at the Census Bureau

by
Daniel W. Gillman
Martin V. Appel
Census Bureau

Abstract

This paper contains the results of continuing research at the Census Bureau into the content, design, population, query, maintenance, and implementation of a statistical metadata repository and the tools to use it. The goals of the research are many, but the ultimate goal is to implement a production statistical metadata repository and its associated tools and services at the agency.

The vision for statistical metadata management at the Census Bureau is to build a metadata repository with Internet/Intranet access to support Internet data dissemination and automated survey design and processing tools. The repository will contain information describing the design, processing, analysis, and data for all the surveys the agency conducts. It will be possible to locate all the available information describing a single survey and locate specific types of information for any set of surveys. This metadata will be available over the lifetime of a survey and its data (which may exceed the life of the survey). Metadata will be collected automatically as a function of automated survey design and processing tools. Metadata describing surveys, variables, data sets, products, and other objects will be obtained through a variety of searching capabilities. Detailed terminologies (e.g. thesauri, taxonomies, or ontologies) will aid the search for specific items or subjects of interest.

In support of the goals, a multidimensional effort was launched. The major parts of this effort include:

- development of detailed models for describing the content and organization of a statistical metadata repository;
- development of a statistical metadata repository and tools for the collection, registration, and query of metadata;
- integration of the repository with other statistical information systems (i.e. Internet data dissemination and automated survey design/production tools); and
- development of terminologies for different subjects, views, and search methods.

The paper will describe the models and terminologies that are developed or under development.

Collecting the metadata to populate the repository is not easy. Survey designers and analysts often create metadata only for themselves and sometimes as an afterthought. When asked about the importance of metadata, the designers and analysts say that it is important. Then, they say they don't have the time or resources to enter it into a repository. Effective, automated survey design and processing tools will collect the metadata without appreciable extra effort. Success is achieved when the users perceive the repository as an indispensable part of their work.

Metadata repository tools are divided into several types: collection, registration, crosswalk, maintenance, query. The paper will describe recent efforts to build automated survey processing tools that provide metadata collection as a byproduct of their functions. Registration, which is a process to ensure and track the quality of metadata, is described.

1.0 Introduction

This paper contains the results of continuing research at the U.S. Bureau of the Census (BOC) into the content, design, population, query, maintenance, and implementation of a statistical metadata repository. The goals of the research are many, but the ultimate goal is to prove the need and feasibility for agency wide statistical metadata management at the BOC.

The vision for statistical metadata management at the Census Bureau is to build a metadata repository with Internet/Intranet access to support Internet data dissemination and automated survey design and processing tools. The repository will contain information describing the design, processing, analysis, and data for all the surveys the

agency conducts. It will be possible to locate all the available information describing a single survey and locate specific types of information for any set of surveys. This metadata will be available over the lifetime of a survey and its data (which may exceed the life of the survey). Metadata will be collected automatically as a function of automated survey design and processing tools. Metadata describing surveys, variables, data sets, products, and other objects will be obtained through a variety of searching capabilities. Detailed thesauri, taxonomies, and ontologies will aid the search for specific items or subjects of interest.

The metadata needs for the BOC were analyzed, and three models were developed which describe statistical surveys and the underlying variables (data elements) which are measured or reported by these surveys. Two of the models, a survey process model (called the Table of Contents or TOC) and a business data model (BDM), were developed to describe and document the design, processing, and analysis for statistical surveys. The information about variables (or data) is captured in a data element model (DEM) built from national and international standards on data elements.

In support of the goals, a multidimensional effort was launched. The major parts of this effort include:

- development of a statistical metadata repository based on the models and tools for the collection, registration, and query of metadata;

- integration of the repository with other statistical information systems (i.e. Internet data dissemination and automated survey design/production tools); and

- development of terminologies for different subjects, views, and search methods.

The paper will describe the models and terminologies that are developed or under development.

Collecting the metadata to populate the repository is not easy. Survey designers and analysts often create metadata only for themselves and sometimes as an afterthought. When asked about the importance of metadata, the designers and analysts say that it is important. Then, they say they don't have the time or resources to enter it into a repository. Effective, automated survey design and processing tools will collect the metadata without appreciable extra effort. Success is achieved when the users perceive the repository as an indispensable part of their work.

Metadata repository tools are divided into several types: collection, registration, crosswalk, maintenance, query. The paper will describe recent efforts to build automated survey processing tools that provide metadata collection as a byproduct of their functions. Registration, which is a process to ensure and track the quality of metadata, is described.

2.0 Definitions

Statistical Metadata 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.

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 planning, design, maintenance, implementation, processing, operation, and evaluation types of activities.

Statistical Metadata Repository is a repository of statistical metadata and pointers to other metadata (such as documents or images). A prototype system is in operation. A pilot production system called the **Corporate Metadata Repository** (CMR) is under development. The CMR is planned to be the statistical metadata repository for the Census Bureau.

The CMR is designed to assist with two new types of tools which are under development at the BOC: Internet data dissemination ; and automated integrated survey processing systems. These systems are known formally as **Statistical Information Systems** (SIS). See figure 1 for an overview of the architecture of a SIS.

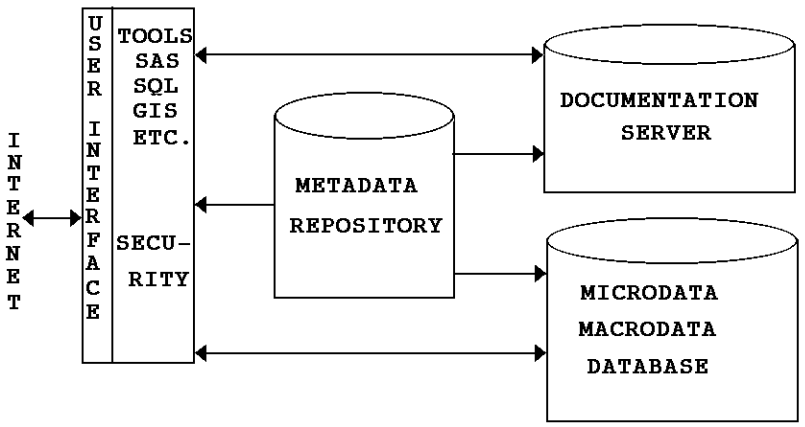

Figure 1. Architecture of Output-Oriented SIS

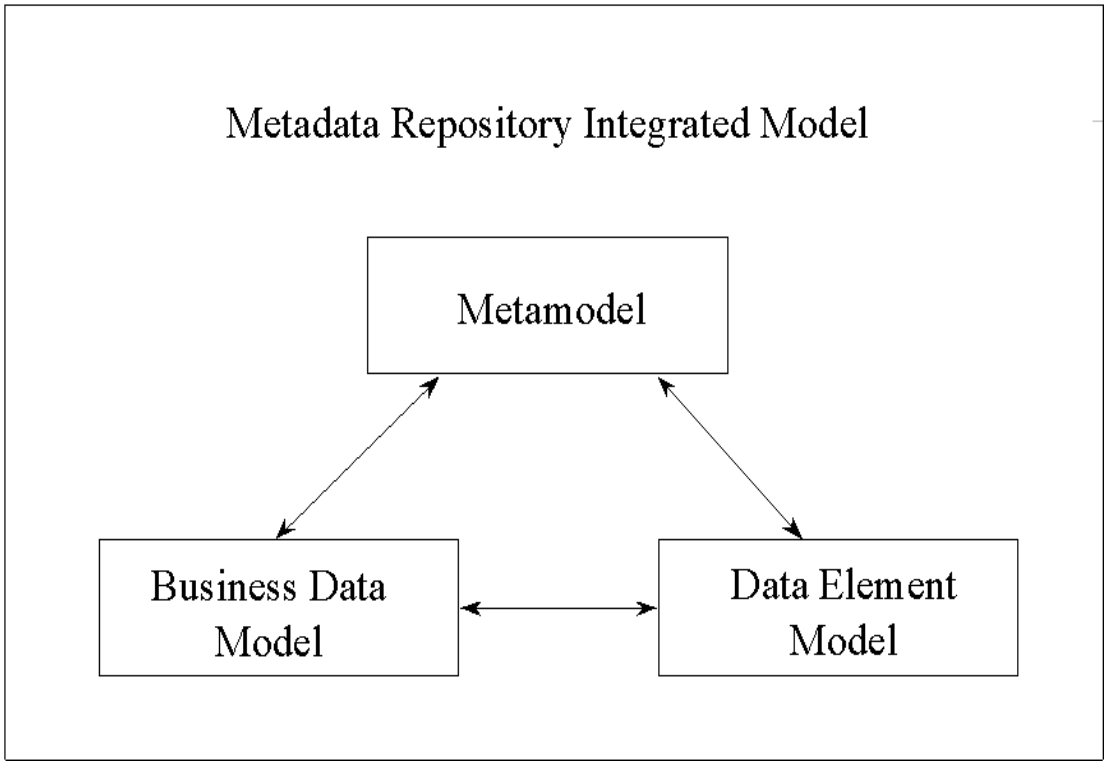

Figure 2: Overview of Integrated Repository Model

3.0 Statistical Metadata Repository

This section describes the design and content of the CMR.

3.1 Models

The design of the CMR is based on three data models. Within the repository, these models were integrated into one extensive model. Extensions to the model are planned as new items or needs are identified.

The three models represent the major dimensions to the CMR model (see figure 2). They are described briefly here and are discussed in more detail below:

- **Business Data Model** - The model describes the business of the BOC - surveys. It describes survey designs, processing, analyses, data sets, products, and documents as related to statistical surveys.
- **Data Element Model** - The data element model is a structure for managing the names, definitions, permissible values, and other attributes of data elements.
- **Metamodel** - This model describes application specific areas and other non-business related items such as security, access control, database schemas, record layouts, and time frames.

The CMR also makes use of a business process model:

- **Table of Contents** - A business process model was also developed. It is in the form of an outline, or table of contents (TOC). The TOC describes the processes of a survey from design to data dissemination.

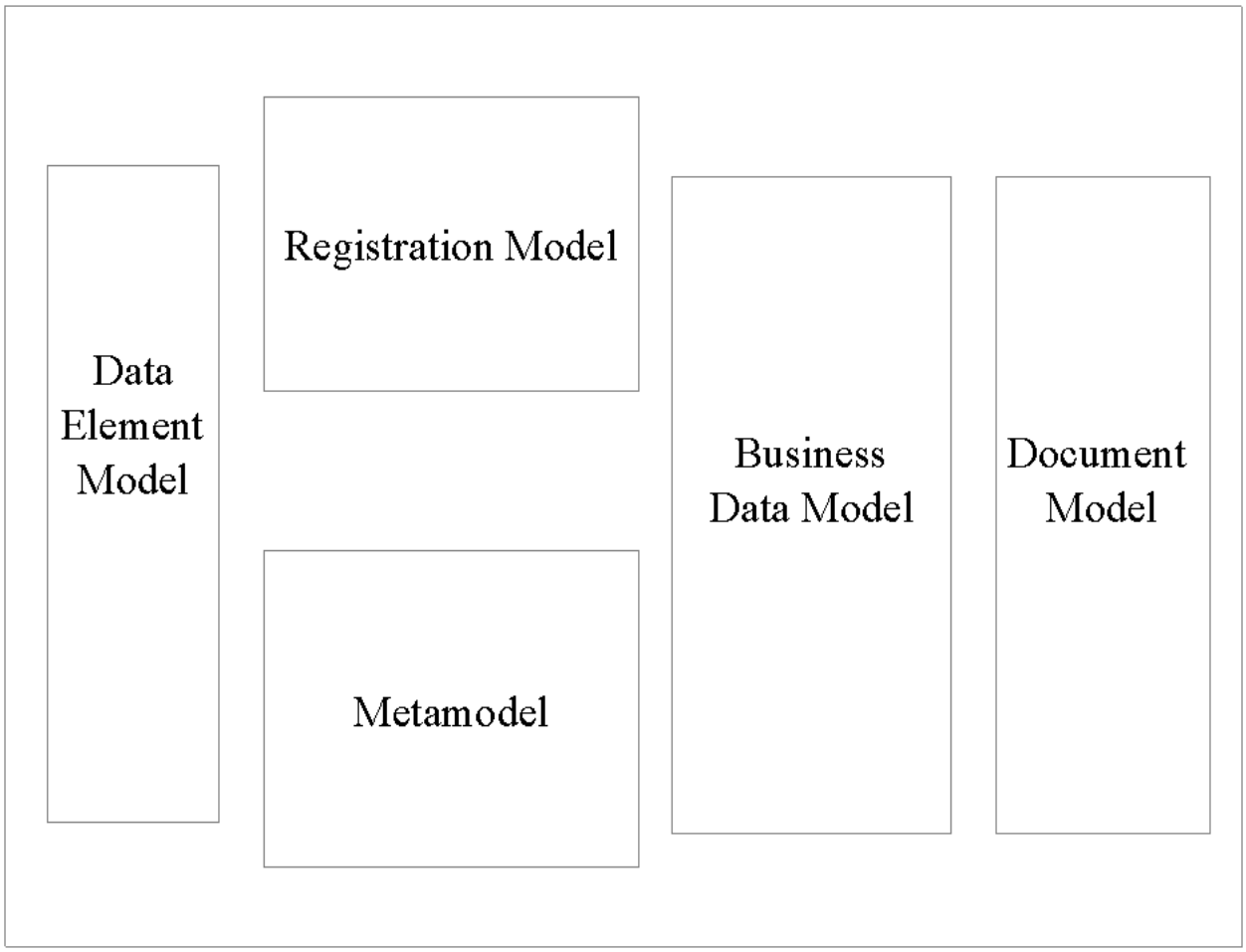

Figure 3: Repository Model Functional Areas

The CMR model is divided into five functional areas. This view provides a clearer picture of how the integrated model works (see figure 3). The functional areas are:

- **Data Element** - Manages the names, definitions, permissible values, and other attributes of data elements.
- **Registration** - Manages the metadata needed to register objects in the repository, i.e. the common information needed to describe each of the objects which are registered in the CMR, much like an electronic card catalog.
- **Metamodel** - Manages the application specific information such as security and access control, search criteria, record layouts, database schemas and access, etc.
- **Business Data** - Manages information about surveys, including design, processing, and analysis.
- **Documentation** - Manages information about documents. The association of documents to different records within other parts of the model acts as part of a classification system for the documents.

3.1.1 Business Data Model[1]

The Business Data Model (BDM) describes the business (statistical surveys) of the BOC. It is composed of entities, attributes, and relationships which describe information that a statistical agency needs to keep about surveys. The model supports the storage of metadata as single attributes or as documents. Figure 4 is a high level ER diagram of the BDM.

The BDM describes survey designs, processing, analyses, and data sets. It contains entities for each of the important parts of a survey: universe, frame, sample, questionnaire, etc. The model allows for the organized storage and search for metadata about a survey, and it allows searching for metadata items across surveys. Many statistical metadata systems in use today address the metadata needs for a single survey or application, but the BDM addresses the metadata needs for many surveys.

The model also provides several other features listed below:

- maintains a list of all current surveys conducted by the agency;
- allows for comparing designs, specifications, or procedures across surveys;
- allows for reuse of designs, specifications, or procedures;
- provides for categorizing and classifying documentation;
- provides for assembling complete documentation for a survey.

3.1.2 Data Element Model[2]

Data elements (or variables) are the fundamental units of data an organization collects, processes, and disseminates. The data element model (DEM) is a structure 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 dictionaries are usually associated with single data sets (files or databases), but a DER contains information about the data elements for an entire program or organization. The information contained in a DER is part of an organization's metadata. Therefore, the registry itself will be part of the CMR.

Important applications for DERs include SISs. 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. Automated integrated survey processing systems that will include sample and questionnaire design, automated edits and imputation, and coding systems require full descriptions of data elements. Designers need to know the definitions of all variables that may be affected by the programs they are using.

The DEM model provides for all the metadata needed to describe data elements. It also provides the entities necessary for registration and standardization of data elements. Generalizing the concept of registration (see section

[1] Designed with the assistance of Database Design Solutions, Inc. of Bernardsville, NJ.
[2] Designed with the assistance of Metadata Management Corporation, Ltd. of Vienna, VA.

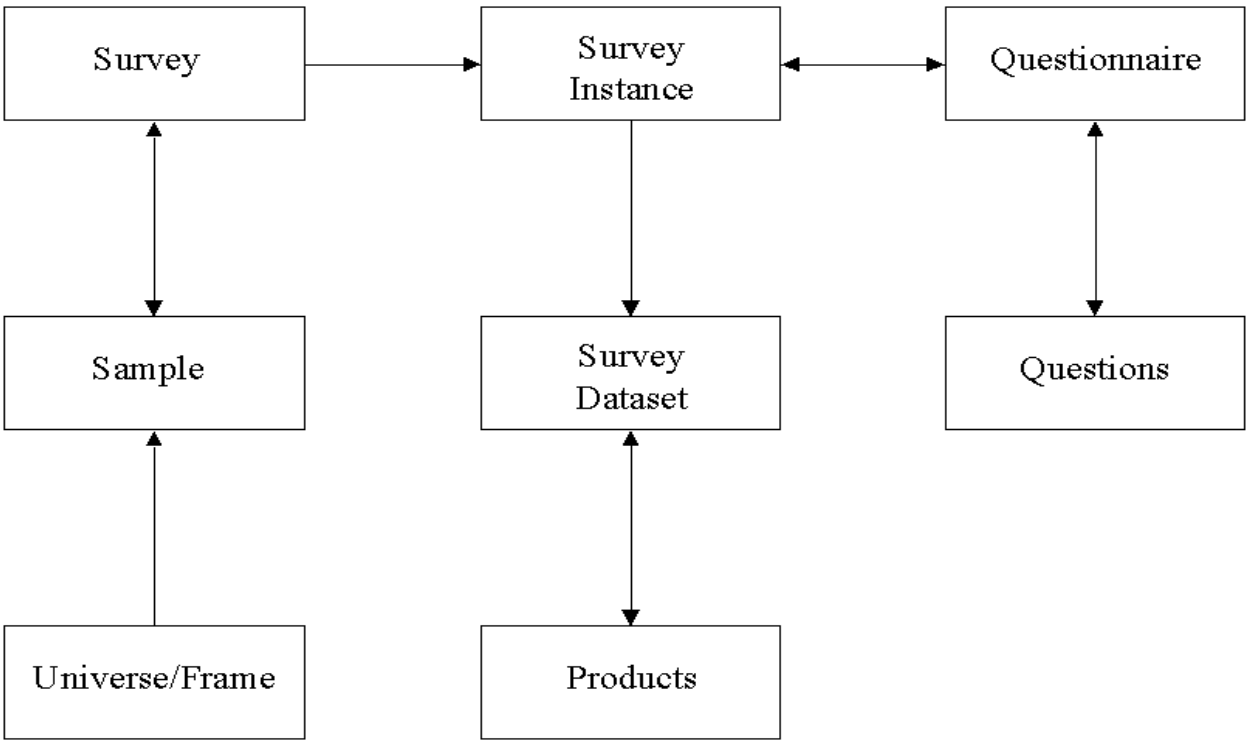

Figure 4: High Level E-R Diagram for Buisness Data Model

4.2 below) to include documents, data sets, products, and surveys provides a framework for merging the DEM and the BDM. A consequence of registering the important metadata items in the CMR is that the repository, from the registration point of view, is somewhat like a card catalog of metadata items. The integration must also include linking data elements to each of the entities in the BDM which use them (e.g. frame, sample, survey data set, question, etc.).

An important feature of the DEM is that data elements are composed of a concept (data element concept) and a representation or value domain (set of permissible values). The power of this is seen as follows:

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

See figure 5 for a high level ER diagram of the DEM.

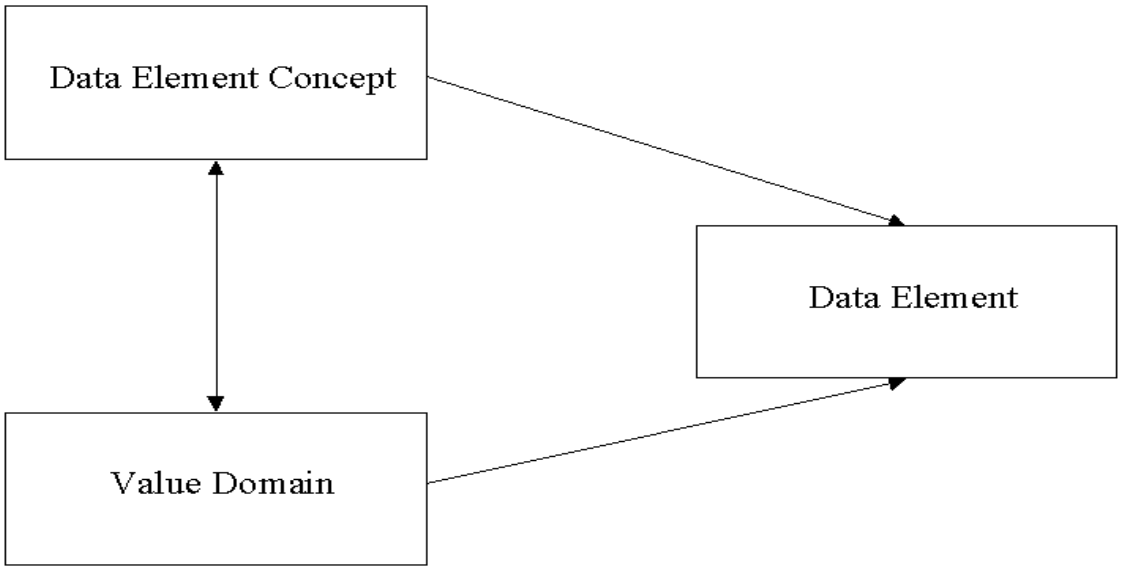

Figure 5: High Level E-R Diagram of Data Element Model

Data elements are composed of three parts as follows[3]:

- the **object class** is a set of ideas, abstractions, or things in the real world that can be identified with explicit boundaries and meaning and whose properties and behavior follow the same rules;
- the **property** is a peculiarity common to all members of an object class; and
- the **representation** describes how the data are represented, i.e. the combination of a value domain, datatype, and, if necessary, a unit of measure or a character set.

Object classes are the things about which we wish to collect and store data. Examples of object classes are cars, persons, households, employees, orders, etc. However, it is important to distinguish the actual object class from its name. Ideas simply expressed in one natural language (English), may be more difficult in another (Chinese), and vice-versa. For example, "women between the ages of 15 and 45 who have had at least one live birth in the last 12 months" is a valid object class not easily named in English. Nevertheless, object classes can be formed by combining two or more other object classes. This example combines the notions of "people between the ages of 15 and 45" with "women who have had live births in the last year".

Properties are what humans use to distinguish or describe objects. Examples of properties are color, model, sex,

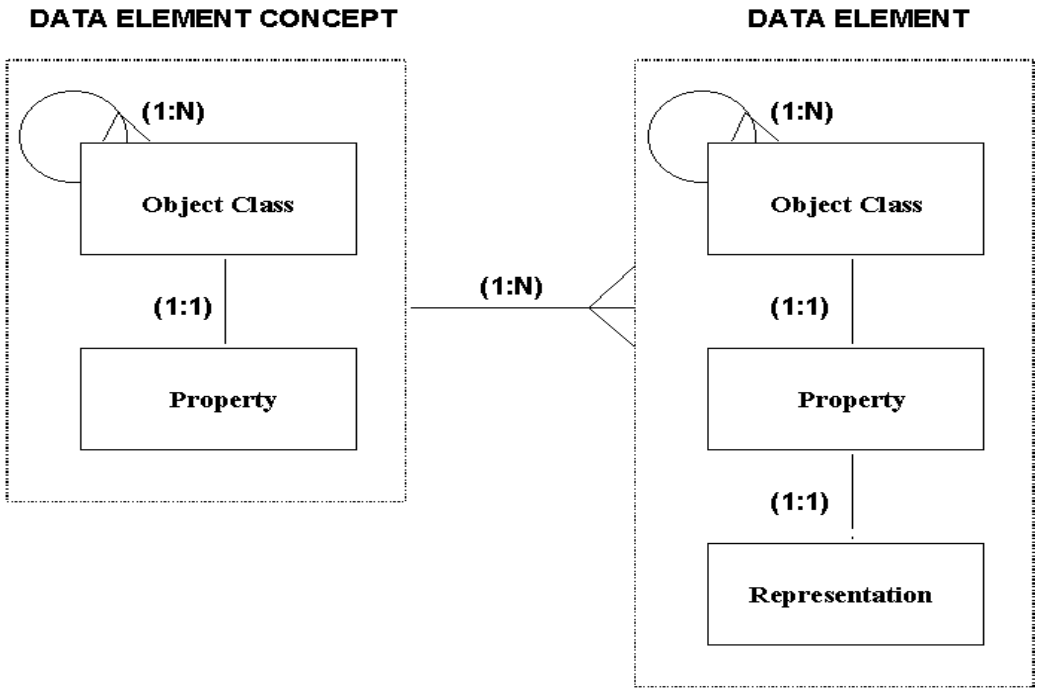

Figure 6: Fundamental Concepts of Data Elements

[3] Taken from Part 1 - Framework - of ISO/IEC 11179, the Specification of Standardization of Data Elements.

age, income, address, price, etc. Again, properties may need to be described using multiple words, depending on the natural language in use.

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

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

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

The combination of an object class and a property is a **data element concept** (DEC). A DEC is a concept that can be represented in the form of a data element, described independently of any particular representation. In the examples above, annual household income actually names a DEC, which has two possible representations associated with it. Therefore, a data element is also composed of two parts: a data element concept and a representation. Figure 6 illustrates the ideas discussed above.

3.1.3 Metamodel

The metamodel is the repository's view of itself. It contains application specific entities necessary for the functioning of particular SISs, and information which controls access to metadata in the rest of the repository. The kinds of information the metamodel handles are access control, security, physical location of data, machine addresses, record layouts, database schemas, access procedures, etc.

The development of the metamodel has been iterative. No specific metamodel has been built. Instead, as new functions are identified, they have been added it. The partnerships (see section 3.3) that have been formed with SIS developers for using the CMR model have been a rich source for new entities and attributes. The metamodel and the CMR model in general have been enhanced because of these partnerships.

3.1.4 Business Process Model

A table of contents (TOC) outline view (see Census Bureau, 1996) of survey processes was developed. It was patterned after work done by a BOC Reinvention Lab and at Statistics Sweden (see Rosen and Sundgren, 1991). The TOC is formally a Business Process Model. It is divided into eight chapters, each detailing a different aspect of survey processing. The chapter names and their descriptions follow below:

- *Content* - The Content refers to the nature of the information that is the subject of the survey, i.e. what the universe is, a description of the data collected, and a description of the resulting products. May contain definitions, and data standardization and coding information.
- *Planning* - Documentation related to the planning and management of the design; the conduct of the survey and the analysis, dissemination and disposition of the data. This includes documentation related to budgeting, manpower, and training.
- *Design* - The design and specifications for how the survey will be conducted. Includes the design of the frame, sample, and questionnaire; and the specifications for edits, coverage, and estimations.
- *Data Collection* - Obtaining information from respondents and the conversion of that data into a form which can be processed.
- *Data Processing* - The stage of a project, following collection and receipt of the original material and preceding report-writing, during which the information is entered onto a machine-readable medium (or directly into a computer system) and eventually used to produce tabulations and statistical analyses.

- *Data Analysis* - Documentation related to all statistical processes used to analyze the survey results or those used for displaying or presenting the resultant information.
- *Data Dissemination* - The process of making data available to users, electronically or otherwise. Electronic data dissemination includes use of the Internet or CD-ROMs.
- *Data* - Any information gathered as the result of a survey or added to a survey form.

There are two uses that are being developed for the TOC: 1) to be used as a check list for users who need to provide metadata or users who want to search metadata from the CMR (see section 4.2); and 2) to serve as a mapping between the CMR and other repositories which need to share metadata (see Gillman, Appel, and LaPlant, 1996). In particular, the TOC can be used as a means to classify documents from another repository in the CMR.

3.1.5 Registration Model

Registration is the process of providing the CMR with its knowledge about the metadata, e.g. name, contacts, stewards, location, type, etc. Some of the general classes of items which need to be registered are data elements, surveys, products, data sets, and documents. Registration requires several things:
- all the necessary attributes are specified;
- all the necessary links are made (e.g. linking a data set to all the data elements in its data dictionary);
- classifying the registered item.

Registration also manages metadata content and quality. It is relatively easy to conceive of entering new data into a database. However, registration includes
- making sure mandatory attributes are filled out;
- determining that rules for naming conventions, forming definitions, classification, etc. are followed;
- maintaining and managing levels of quality.

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

Registration is the process that turns the CMR into a library of information. The common metadata about objects the CMR describes are captured. Figure 7 provides an overview of the registration model.

3.1.6 Documentation Model

Documents are a very important source of metadata in the BOC. Designs, specifications, procedures, and manuals are all types of documents. The CMR model and the TOC are the underlying categorizing and classification system for documents within the statistical survey business framework.

The CMR model also contains a means for managing the metadata that describes documents: the Documentation section of the CMR model.

The card catalog metaphor described above is used to advantage for modeling documents. If documents exist which describe some object, then those documents are attached to the object at registration and not when the object is placed in its proper place within the BDM or DEM. The disadvantage to this approach is the lack of a specified document set in the CMR for every survey the BOC conducts. The current CMR model supports *ad hoc* document sets, i.e. each survey management group defines what kind of documents will be produced to describe its survey.

Figure 8 depicts an overview of the Documentation Model portion of the CMR model.

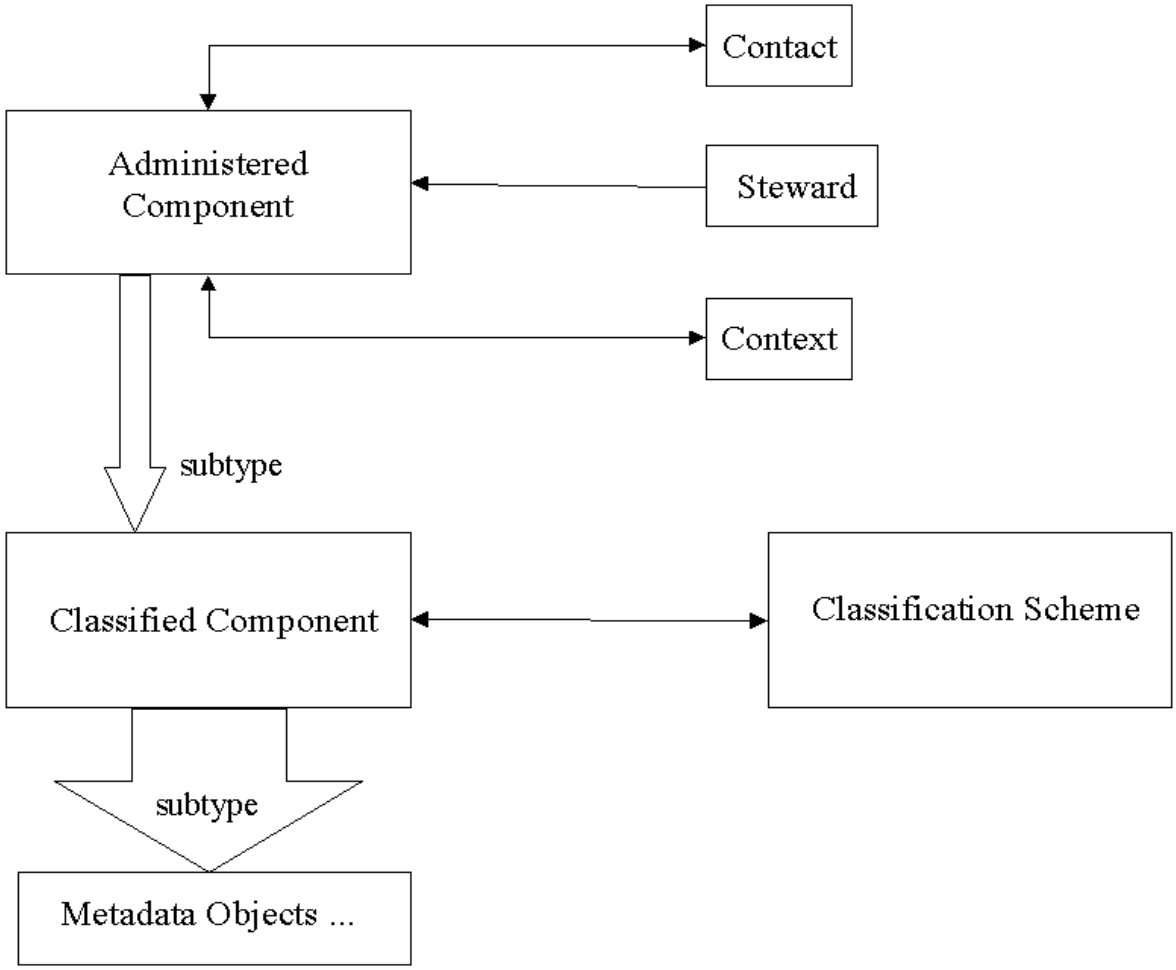

Figure 7: Overview of Registration Model

3.2 Standards

In this section the applicable standards which have been used to guide the development of the CMR and its associated tools will be described briefly. Figure 7 shows how the standards discussed below fit into the design of the CMR.

3.2.1 Data Element Standards

The model for the data element portion of CMR is based on the conceptual framework contained in the ANSI standard, **The Metamodel for the Management of Shareable Data (MMSD)**, ANSI X3.285. It, in turn, incorporates all the principles described in an international standard, **Specification and Standardization of Data**

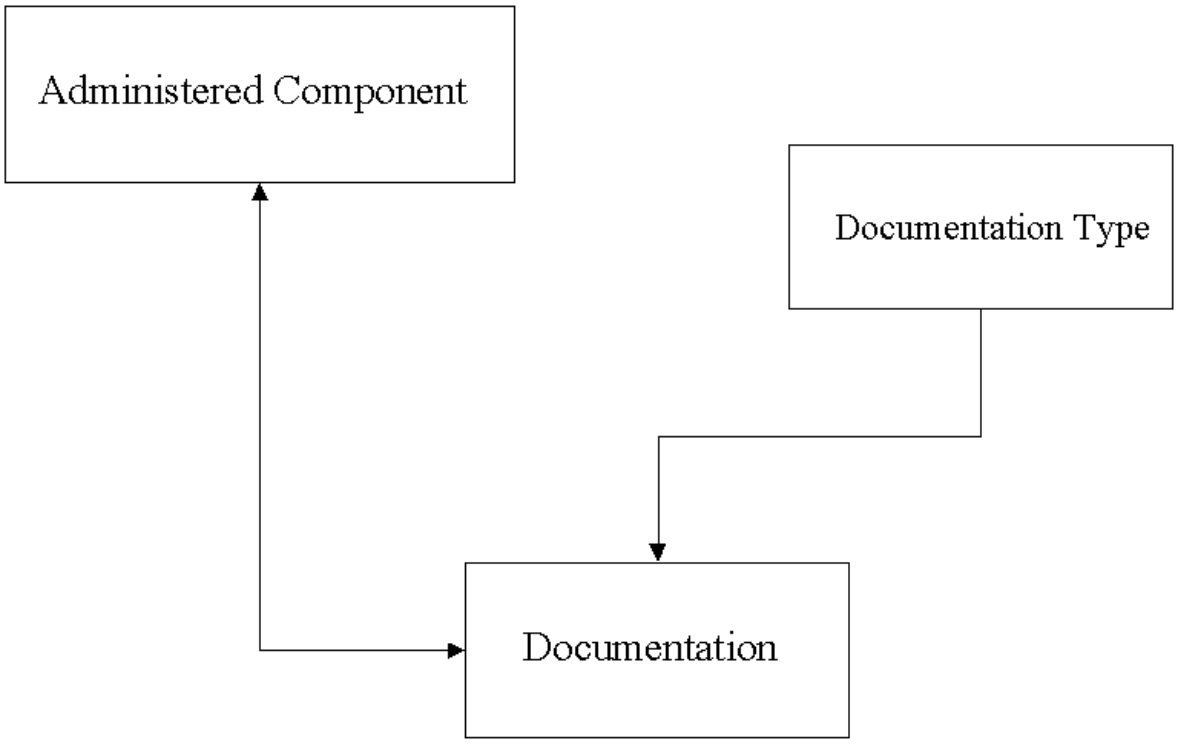

Figure 8: Overview of Documentation Model

Elements , ISO/IEC 11179 (see ANSI X3L8, 1996). ANSI X3.285 provides a conceptual model for building a data element registry and contains some extensions to the framework described in ISO/IEC 11179.

The MMSD metamodel provides a detailed description of the types of information which 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 provides users the information they need to understand the data elements of an organization.

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

- Part 1 - *Framework for the Specification and Standardization of Data Elements* - Provides an overview data elements and the concepts used in the rest of the standard. The current status of this document is *Final Draft International Standard.* One of the authors is the editor of this part.
- Part 2 - *Classification of Data Elements* - Describes how to classify data elements. The current status of this document is *Final Draft International Standard*.
- Part 3 - *Basic Attributes of Data Elements* - Defines the basic set of metadata for describing a data elements. This document is an *International Standard*, and it is under revision. The revision will include ANSI X3.285.
- Part 4 - *Rules and Guidelines for the Formulation of Data Definitions* - Specifies rules and guidelines for building definitions of data elements. This document is an *International Standard.*
- Part 5 - *Naming and Identification Principles for Data Elements* - Specifies rules and guidelines for naming and designing non-intelligent identifiers for data elements. This document is an *International Standard.*
- Part 6 - *Registration of Data Elements* - Describes the functions and rules that govern a data element registration authority. This document is an *International Standard.*

3.2.2 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, 1997) is a draft statistical metadata content standard for the BOC, undergoing formal review. It provides a description of the information or documentation about statistical data. The content and design of the standard is based primarily on the BDM. The entities of the BDM specify the content sections of the SDSM.

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.

Each of the more than 30 sections in the SDSM consists of a list of entries, some that reference other sections. Each entry is a metadata data element. Any of these metadata data elements may be used to identify specific instances of metadata. The metadata may be some specific information (such as a number or text) or a **url** to a file of some type (e.g. documents, gif's, etc.).

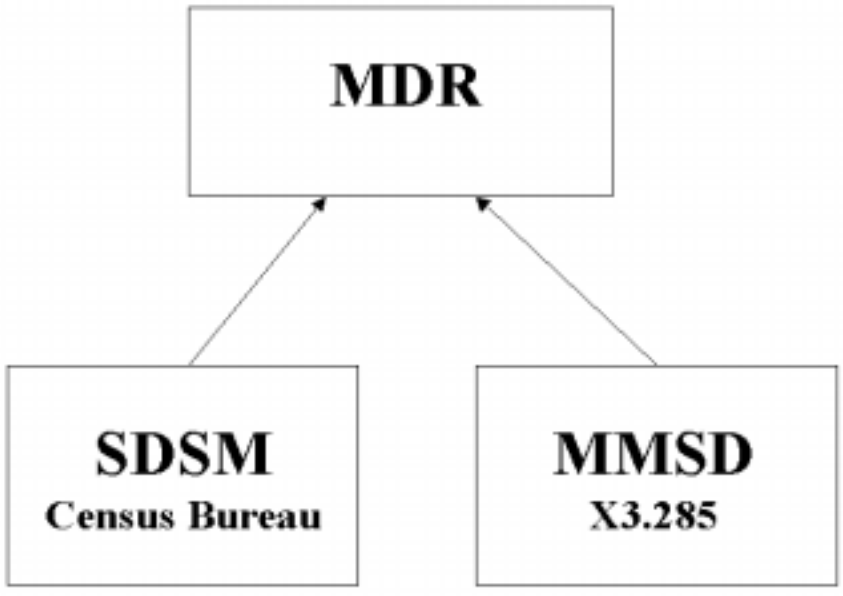

Figure 9: Integration of Standards in CMR Design

4.0 Tools

The development of useful tools for metadata collection, query, registration, classification, and quality management will make the CMR useful for the BOC in a production environment for a long time. Query tools are much like those needed for any databse application, so won't be discussed. Registration, classification, and metadata quality management are closely related, so are discussed together.

4.1 Collection and Management

Collecting metadata is recognized as a very difficult problem. Survey units at the BOC create and manage metadata in their own ways. The idea of storing the metadata that is already created in a "centralized" repository is seen as a lot of extra work. A way around this problem is to build automated survey design and processing tools that collect metadata as a by-product of their use.

These and other tools might are being developed as the CMR production system moves forward. Careful analysis of the culture at the BOC and interviews with subject matter experts around the agency show that this approach has the highest probability of success.

Reasearch centers on new ways of managing and modeling metadata, such as with Object Oriented Database Management Systems and UML (Unified Modeling Language), defining common interfaces to increase interoperability (e.g. using CORBA (Common Object Request Broker Architecture)), and building transfer mechanisms such as with XML (Extended Markup Language).

4.2 Registration, Classification, and Quality Management

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

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

Metadata quality is measured by the ability of the user to search, locate, and understand objects the metadata describe. There are several factors that contribute to this quality:

- The set of metadata (descriptive) attributes is "complete";
- The mandatory metadata attributes describe each object uniquely;
- The rules and guidelines for providing data in those attributes are well-defined;
- Measures determining how well the rules are followed are specified;
- Rules and guidelines for classifying objects are well-defined;
- Measures for determining how well an object is classified are specified.

Research will focus on ways of measuring quality using these and other criteria which are found to be important.

The registration process is where the quality of metadata is monitored. The Registration Authority, the business unit in charge of managing the CMR and the registration process, is responsible for maintaining (measuring) quality. The steward (or subject matter expert) is responsible for providing the metadata by the rules.

One quick way for users to learn the quality of metadata describing an object of interest is through the use of a *registration status*. This status has several levels specified in advance, and the Registration Authority assigns a level to an object dpending on the quality of the metadata that is currently in the repository. The status changes as the quality of the metadata changes.

There is also a human side to metadata registration which must be developed. Metadata and data administration is required to manage the content of the metadata that is registered. No function of this type exists at the BOC at this time at the agency level. Some of the necessary functions are:

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

Metadata administration will require a large commitment from the BOC, but it will greatly enhance the usefulness of BOC data, make the CMR a better tool, and facilitate the sharing and understanding of data and metadata among groups within the BOC or with other agencies.

The development of useful, semantically rich, and accurate classification schemes is a major task facing statistical agencies. However, people have begun this work and met with some success. The medical community has vast experience in this area.

Classification is a very important subject because correctly classified objects are much easier for users to locate. In general, a classification scheme is an organized set of terms. Several types of classification schemes are terminologies, taxonomies, thesauri, and ontologies. Researchers at Columbia University and University of Southern California Information Systems Institute are building an ontology to describe energy data, especially around the term "gasoline". Discussions at inter-agency meetings quickly showed that this term referred to data collected by many different agencies.

Classification is the process of assigning terms of a terminology (e.g. thesaurus, taxonomy, or ontology) to an object. The terms have pre-specified meanings, and subject matter experts assign all the relevenat terms which describe an object. The result is a description in the form of a set of terms which describes an object to a user. Both meaning and a context for that meaning are possible to convey with a good terminology. For example, the term "sample" has a similar but different meaning in statistical survey work as compared to scientific measurment work. If the user knows the context, the term has more meaning.

Performing this task well will enable users to locate and understand objects (data) quickly and accurately, and users will be able to discern small differences in meaning. Developing and applying and measures of semantic difference will further enable users to see differences. For users trying to harmonize data, the semantic difference measures are crucial to uncovering possible candidates for harmonization.

Useful classification schemes already exist which can be incorporated into registration tools, such as the TOC (described above) and the **themes** as specified in the **Cultural and Demographic Data Metadata** draft standard of the **Federal Geographic Data Committee**.

It will be useful for the BOC to build a taxonomy of statistical terms to help with the classification problem. Of course, effective classification schemes also help with the search for metadata and for understanding the semantics of data or metadata.

5.0 Implementation

The CMR is not an end in itself, but an integral part of SIS's within the BOC. FERRET (Federal Electronic Research and Retrieval Extraction Tool) and AFF (American Fact Finder) are two major Internet data dissemination systems under development at the BOC. Both require a metadata repository to make them work (see Figure 1). These repositories are considered part of the CMR or applications that communicate directly with the CMR.

Document Management Systems (DMS) have an important role within the BOC. Currently, each subject matter area manages documents in its own way. A DMS will standardize the way documents are managed throughout the BOC. Many types of documents are managed by a DMS; some are important as statistical metadata. Section 3.1.6 (Documentation Model) is a description of how documents which are statistical metadata are handled, categorized, and classified in the CMR.

The pilot projects associated with implementing the intial CMR at the BOC focus on collection tools. A partial list of the possible tools and their purposes is:

- Batch and interactive CRUD[4] tools for managing the CMR and inputting legacy metadata;
- Interface with automated tool for designing electronic questionnaires;
- Linking data sets from different sources (e.g. demographic and economic data) allowing management and comparisons of data elements and survey designs;
- Automatically registering and classifying documents through a document management system.

In order for the pilot projects to succeed, the developers first focused on obtaining support from users and designing the architecture for the system before actual system building started. These items are critical for success, but they take a long time to complete. The initial work took a year. Now, the work is identifying and building pilot projects (see above) with the Economic Directorate at the BOC. As more projects are begun and demonstrated, more users will be willing to try pilot projects of their own. In time, the CMR will be an integral part of the operations of the BOC.

6.0 Conclusion

This paper has discussed the work at the BOC to design a statistical metadata repository using detailed models developed in conjunction or as an implementation of standards developed by international, national, and U.S. Government organizations. The detailed data and metadata models have been built and integrated. The integrated model is the basis for the CMR architecture. It provides a structure for storing the metadata which describes survey designs, processing, analyses, and data sets. The model supports a card catalog metaphor for organizing the BOC metadata.

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

A transition plan to move the management of metadata out of the research arena and into a production mode was written. Its plan for creating a special metadata group within the BOC was accepted. At present, an *ad hoc* group is building the first parts of the CMR architecture and functions. The first production CMR system is planned for the end of FY 2000.

However, several large important areas of research still need exploration. The most important areas are:
- translation of CMR model into object model and implementation of the repository in an object oriented database;

---

[4] CRUD = Create, Replace, Update, and Delete

- building a comprehensive thesurus or taxonomy of statistical terms for searching and classification;
- formalizing the specification for statisitical metadata in certain domains such as sample design and incorporating the formalism into the CMR model.

We recognize that these research areas are very complex. Solutions to these problems are not anticipated soon.

7.0 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 (1999), "ISO/IEC 11179 Part 1 - Framework for the Specification and Standardization of Data Elements", Final Draft International Standard, May 1999.
- Census Bureau (1997), "Statistical Design and Survey Methodology Metadata Content Standard", Draft, Census Bureau Internal Document, April, 1997.
- Census Bureau (1996), "Table of Contents for Statistical Design and Survey Methodology Metadata Content Standard", Draft, Census Bureau Internal Document, July 2, 1996.
- 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.
- Gillman, D. W., Appel, M. V., and Highsmith, S. N. Jr. (1997), "Building a Statistical Metadata Repository", Second IEEE Conference on Metadata, Silver Spring, MD, September 16-17, 1997.
- Gillman, D. W., Appel, M. V., and Highsmith, S. N. Jr. (1998), "Building a Statistical Metadata Repository at the U.S. Bureau of the Census", Presented at the UN/ECE METIS Working Group Meeting, Geneva Switzerland, February 18-20, 1998.
- Graves, R. B. and Gillman, D. W. (1996), "Standards for Management of Statistical Metadata: A Framework for Collaboration", ISIS-96, Bratislava, Slovakia, May 21-24, 1996.
- 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.
- 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.
- 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.