The Statistical Metadata Repository: an electronic catalog of survey descriptions at the U.S. census bureau

*by Daniel W. Gillman and Martin V. Appel**

1. Introduction

The U.S. Census Bureau (BOC) is developing a prototype statistical metadata repository for use with Internet data dissemination and automated integrated survey processing tools. The repository will be an electronic catalog of information about survey designs, processing, analyses, and data sets. Access will be through the Internet and the World Wide Web. Substantial background work was done before work to build the prototype could begin.

Statistical metadata is the information and documentation needed to describe and use statistical data sets for the lifetime of the data. The efficient, effective, electronic management of metadata greatly increases the usefulness of those data sets, especially for Internet data dissemination. Statistical metadata can also be used to facilitate survey design, processing, management, and analysis. Automated integrated survey processing systems, which create and use this information, will allow statistical agencies to conduct their programs in ways that were not possible before.

The repository is being designed based on standards and data models. It is being implemented as a relational database and organized through these standards and models. International, American, and internal Census Bureau standards are all being brought to bear in the development of the repository. Three models have been developed and integrated to form the structure of the repository. The models are the Business Data Model, the Data Element Registry Model, and a Metamodel.

Tools for the collection of the metadata and querying the repository are under development. Without the cooperation of the survey designers and analysts who create the metadata, the repository will never be populated. General, intuitive, and easy to use tools must be developed to collect the data. Conversely, the information in the repository will not be useful if it cannot be retrieved in an easy way. A survey Business Process Model, or table of contents, has been developed for users and analysts to find the type of information they may want to provide. This table of contents is being used as a template in the design of the tools. Also, it can be used to design a low level interface for other systems to access and use the repository.

Substantial benefits should be available to the Census Bureau when the repository is functional. It organizes the documents, data sets, and variable descriptions of the agency. The repository will allow for comparisons across surveys (data or designs) which previously have not been easily available. Finally, the repository will make the public information of the agency fully available from a common source. If other statistical agencies around the world adopt similar approaches, the concept of a "single world-wide statistical agency" on the Internet could become reality.

This paper will define what statistical metadata is, describe the design of the repository (including the standards and models), describe the tools under development for populating and querying the repository (including the table of contents outline), and discuss the ramifications for the agency of implementing the repository.

2. 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. 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,

v 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

v 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 a planned repository of statistical metadata and pointers to other metadata (such as documents or images). A proof-of-concept system has been built (see Gillman and Appel, 1994), and a series of prototypes are under development. The design, uses, and functionality of the MDR will be discussed in more detail below.

3. 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 ; and automated integrated survey processing systems . 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).

3.1 Purposes

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

This has led to the management of data in a decentralized and non-uniform way. On one hand, there is a need for the survey management to process and manage their data in the most efficient way. 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 survey data and metadata. 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 them.

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, and it will be discussed in more detail in section 3.2.2.

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 to which external end-users cannot have access. 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.

3.2 Models

The design of the MDR is based on three data models. Within the repository, these models have been integrated into one extensive model which covers many aspects of statistical metadata. Extensions to the model are planned as new items or needs are identified.

The three models represent the major dimensions to the MDR model (see figure 1). They are described briefly here and will be discussed in more detail below:

- Business Data Model - The model describes the business of the BOC - surveys. It describes survey designs, processing, analyses, datasets, products, and documents as related to statistical surveys.
- Data Element Registry Model - A data element registry is a mechanism for managing the names, definitions, permissible values, and other attributes of data elements. Metadata describing data elements is entered into the registry by a process called registration. Expanding the concept of registration to include surveys, products, datasets, and documents, this model handles the needs of registering metadata.
- 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 metamodel provides the repository's view to itself.

The MDR prototype also uses a business process model described below.

- Table of Contents - A business process model has also been 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.

The MDR model can also be divided into five functional areas. This view gives a clearer picture of how the integrated model works (see figure 2).

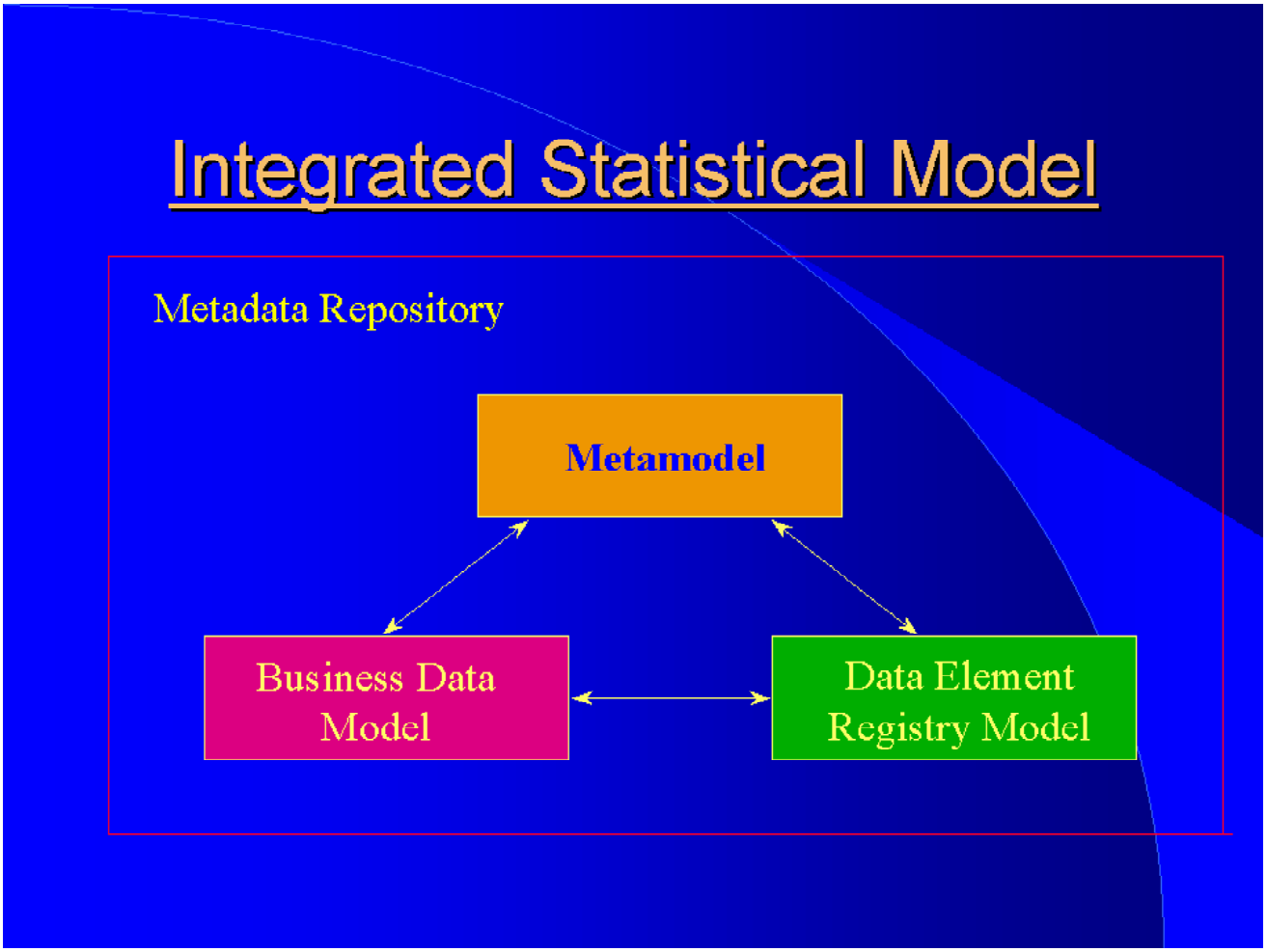

Figure 1: Overview of Integrated Model

The functional areas are:

- Data Element Registry - Manages the names, definitions, permissible values, and other attributes of data elements (see above and below).

- Registration - Manages the metadata needed to register items for which the repository keeps track: surveys, data elements, documents, datasets, products. This section handles the information types which are common to each of the objects which are registered in the MDR, much like an electronic card catalog system.

- Metamodel - Manages the application specific information such as security and access control, search criteria, record layouts, database schemas and access, etc (see above and below).

- Business Data - Manages information about surveys, including design, processing, and data (see above and below).

- Documentation - Manages information about documents. The association of documents to different records within other parts of the model acts as a classification system for the documents.

3.2.1 Business Data Model

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. Much of this information is in the form of specifications or procedural documentation. The model supports the storage of metadata as single attributes or as documents. Figure 3 is a high level ER diagram of the BDM, and see Appendix

A for an entity definition list.

The BDM describes survey designs, processing, analyses, and datasets. 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.

An important feature of the BDM is that documentation is handled in a general way. Each entity of the model allows for many documents to be attached to a single record. The documents can be distinguished by version, document type (e.g. specification, procedure, memo, etc.), the entity the document is associated with, and the relationships the given record has with other records in the model. This provides a comprehensive classification scheme for documents which helps users search directly for the information they need. Coupled with the indexed and key word search provided by most Internet search engines, the BDM is a powerful document management paradigm.

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 assembling complete documentation for a survey.

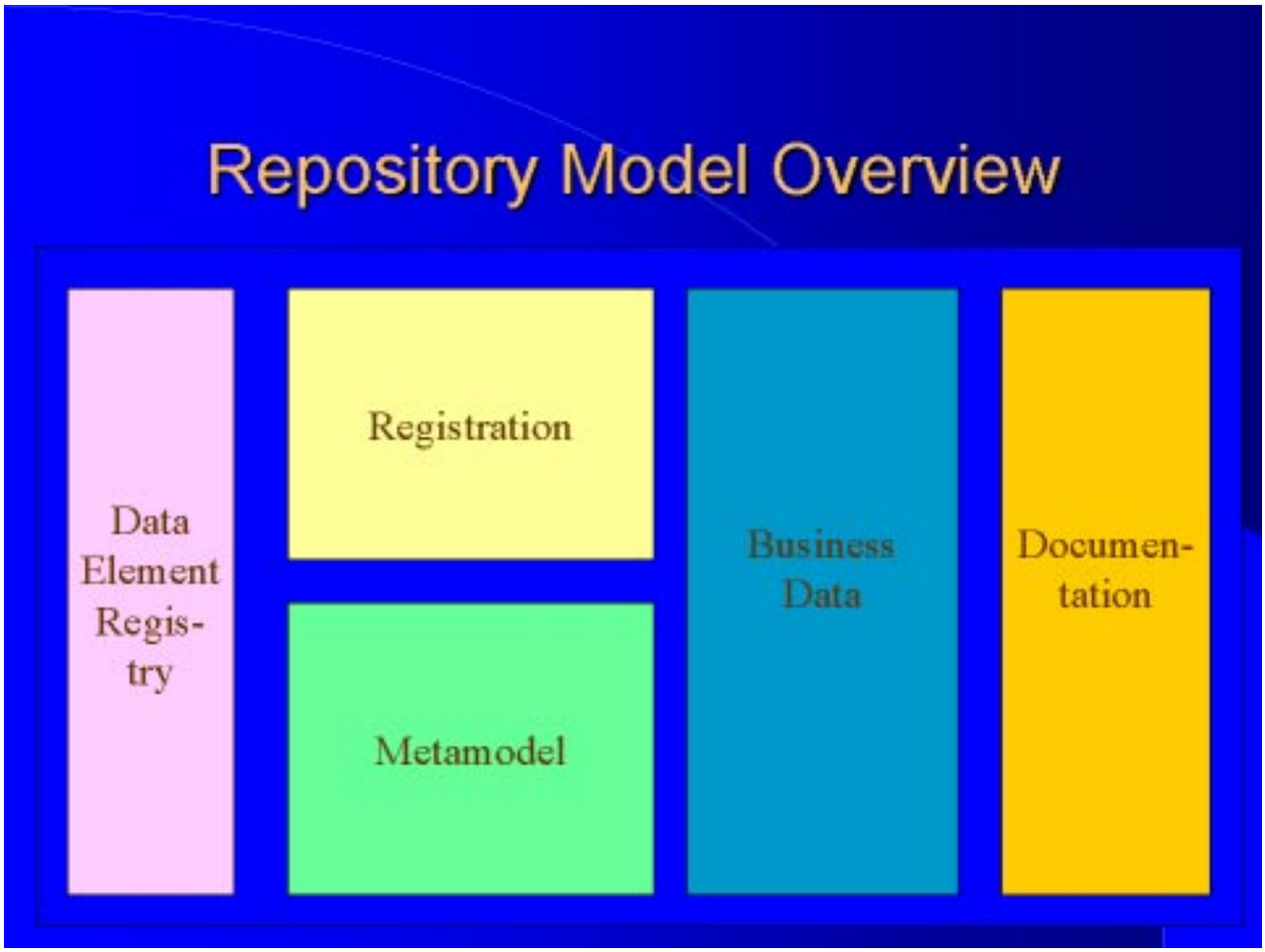

Figure 2: Repository Model Overview: Functional Areas

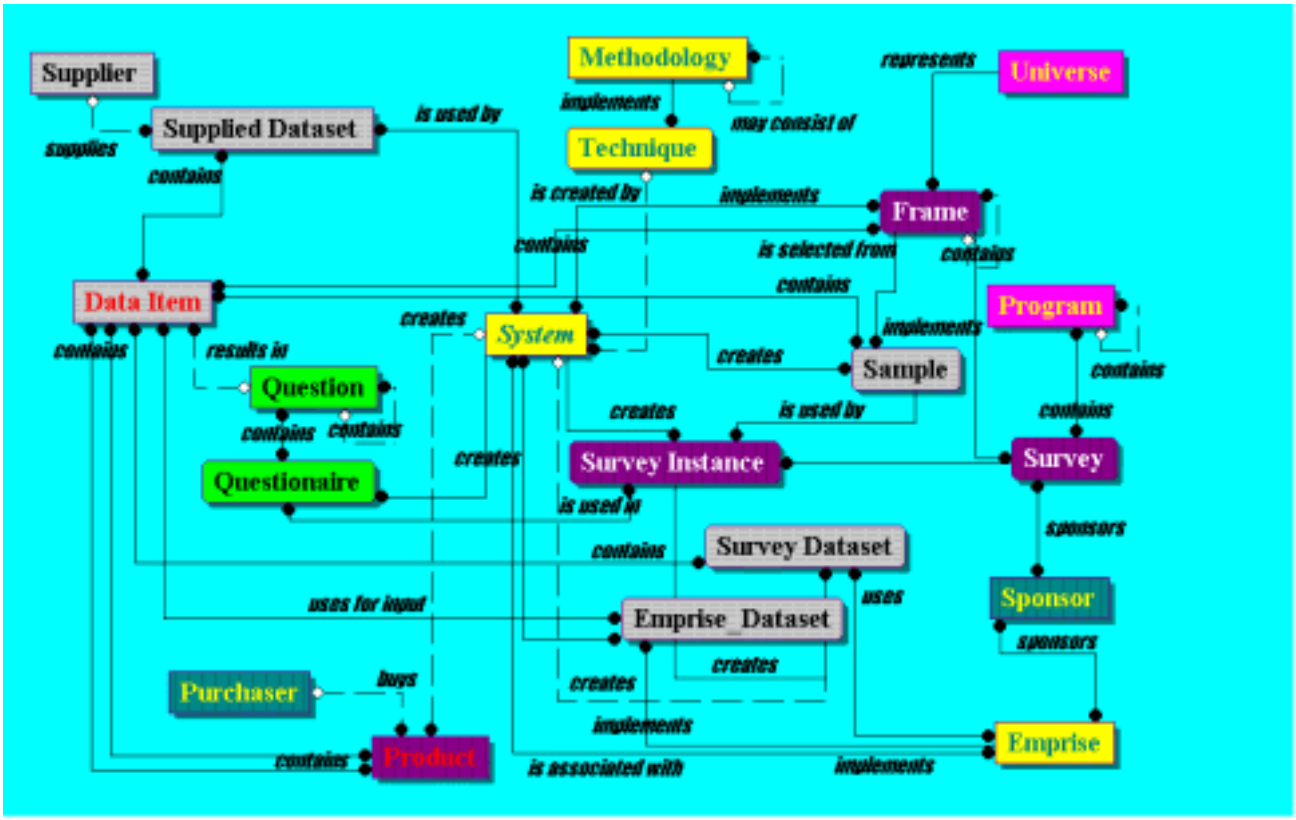

Figure 3: Business Data Model

3.2.2 Data Element Registry Model

Data elements (or variables) are the fundamental units of data an organization collects, processes, and disseminates. A data element registry (DER) is a mechanism for managing data elements in a logical fashion. DER's organize information about data elements, provide access to the information, facilitate standardization, help identify duplicates, and facilitate data sharing. DER's 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 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 MDR.

Important applications for DER's 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. 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 DER 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 4.2 below) to include documents, datasets, products, and surveys provides a framework for merging the DER and the BDM. A consequence of registering the important metadata items in the MDR is that the repository, from the registration point of view, becomes 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 dataset, question, etc.).

An important feature of the DER 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 4 for a high level ER diagram of the DER, and see Appendix B for an entity definition list.

3.2.3 Metamodel
The metamodel is the repository's view of itself. It contains application specific entities necessary for the functioning of particular SIS's, 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 to the MDR model. The partnerships (see section 3.4) that have been formed with SIS developers for using the MDR model have been a rich source for metamodel entities and attributes. As these partnerships continue and the SIS's are further developed, more information is added to the metamodel and to the MDR model.

.3.2.4 Business Process Model
A table of contents (TOC) outline view (see Census Bureau, 1996) of survey processes has been 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

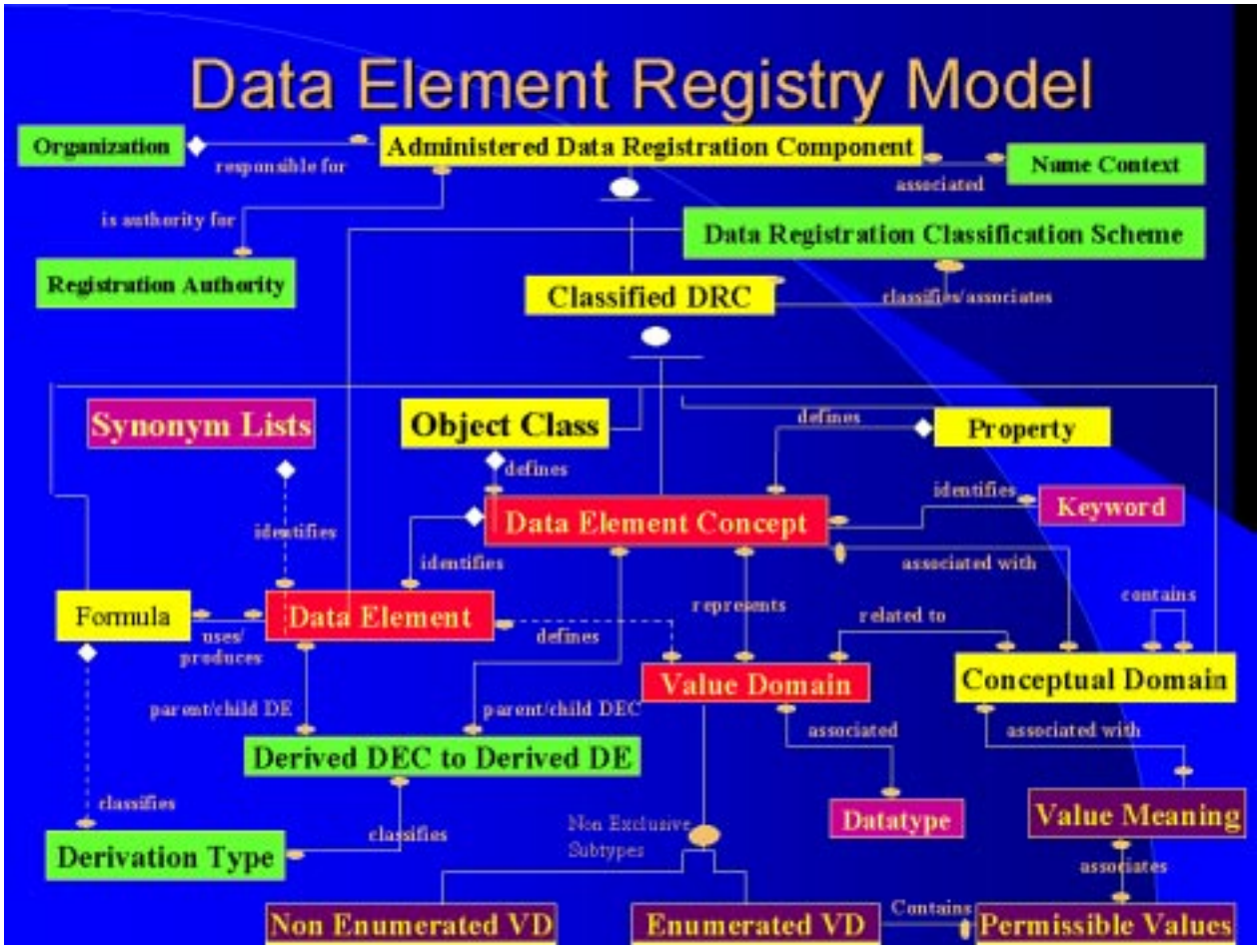

Figure 4: Data Element Registry Model

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 MDR (see section 4.2); and 2) to serve as a mapping between the MDR 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 MDR.

3.3 Standards

In this section the applicable standards which have been used to guide the development of the MDR and its associated tools will be described briefly.

3.3.1 Data Element Standards

The model for the data element registry portion of MDR is based on the conceptual framework contained in the ANSI draft standard, The Metamodel for the Management of Shareable Data (MMSD), ANSI X3.285. 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). 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. A complete data dictionary describing all the entities, attributes, and relationships of the conceptual metamodel is included in this document.

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.

ISO/IEC 11179 is being developed in 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 of the concepts in the rest of the standard. The current status of this document is Committee Draft.

- Part 2 - Classification of Data Elements - Describes how to classify data elements. The current status of this document is Working Draft.

- Part 3 - Basic Attributes of Data Elements - Defines the basic set of metadata for describing a data elements. This document is an International Standard.

- 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.

3.3.2 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. It will provide 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 29 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.)

The SDSM has been submitted to the formal standards review process of the BOC, and is expected to be issued as a BOC standard in Summer 1997. Once this occurs, it is hoped that other statistical agencies will adopt the SDSM or similar standards.

3.3.3 Other Standards
Information Resource Dictionary System (IRDS) is a standard which addresses the use, control, organization, 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). The organization of the MDR model is based on the organization specified in IRDS. See Graves and Gillman (1996) for a more detailed discussion.

The Federal Geographic Data Committee (FGDC) of the U.S. Government has developed a family of metadata standards which addresses the geographic content of data. Executive Order 12906 has mandated that all U.S. agencies that produce geographic based data use these FGDC standards. Most BOC data is based on geography, therefore these standards will apply to BOC data.

Government Information Locator Service (GILS) (FIPS-192) is an extensible standard which describes a format and specifies the underlying protocol (NISO Z39.50) for making metadata available on the Internet. Another Executive Order (the Paperwork Reduction Act of 1995) mandates that all U.S. agencies create and maintain GILS records. This provides a mechanaism for the public to find information about what their government is doing and producing through electronic means.

3.4 Partnerships
Several groups within the BOC developing SIS's have agreed to use the MDR structure to support the underlying metadata needs of those systems. A short description of each SIS follows below.

DADS (Data Access and Dissemination System) is the name for the Census Bureau 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 Census Bureau data. The system will be designed to be fast, flexible, and cost-efficient. To achieve this, four cross-directorate teams were formed to study and recommend policies or designs for user input, promotion and outreach, pricing for products, copyright or trademark, corporate look and feel, data archiving, metadata and documentation, and coordination of various efforts and activities. DADS will attempt to incorporate other work, such as FERRET, where it is appropriate.

The DADS team is following a schedule to produce a new prototype each year, in the month of September, until the full production system is built in 2002. The 1996 prototype was a success, though limited in scope. The MDR model will be used to organize the metadata for the 1997 prototype.

FERRET (Federal Electronic Research and Review Extraction Tool) (see Capps, 1995) is a 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 items (individual survey question items) which can be used to create custom data queries. In addition, users can select macrodata (aggregated or summarized) tables to get preformatted survey data. Results of data queries can be output in SAS datasets or ASCII files. These results can be viewed on the screen or can be downloaded to a local computer. The SAS output allows one to get the results in pie charts, bar charts, or summarized on a U.S. map. The ASCII output can be brought into an Excel spreadsheet.

The FERRET system can be divided into four major parts. The first part is the user interface which is via the World Wide Web. The Ferret repository contains metadata such

as basic variable definitions, keywords, concepts, and other items. The Document Management System handles the documents which describe the survey design, processing, and analysis. Finally, there are two databases handling all the microdata and macrodata.

FERRET currently handles Current Population Survey data. Plans are to add other demographic survey data in the future. Work is also underway to make the FERRET repository model and the MDR model compatible. This will enable people to work with DADS and FERRET systems seamlessly.

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 providing the following basic survey processing functions:

- Data review and correction;
- Edits;
- Imputation;
- Outliers;
- Estimation;
- Estimation variance;
- Disclosure analysis;
- Time series;
- Queries (canned/ad hoc);
- Tables (canned/ad hoc);
- Management information; and
- Survey control operations (for scheduling of batch mode processes).

It will also provide the following additional functions:

- Generate standard and non-standard mail files for mail-out operations;
- Generate standard telephone files for telephone follow-up operations;
- Maintain standard variable names and flags;
- Maintain standard data structures;
- Allow entry of survey design specifications including edit and imputation parameters as determined by analysts or through automated historical data analysis
- Provide audit trails and backup capabilities;
- Provide access to SSEL; and
- Provide access to other economic area surveys and censuses.

The above provides a view of the functionality which StEPS will be designed to provide. Implementation details are not yet available. The StEPS system developers plan to use the MDR as a source of information about variables.

Product Registration is a multi-divisional effort to unify the systems that manage the production, inventory, distribution, and sale of Census Bureau products. The MDR model will be used to register products, i.e. link products to the variables, surveys, geography, and other items that will enable users to locate them. This work has recently started.

3. Metadata Management

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

4.1 Content and Storage

Content refers to the identification of which metadata will be collected and stored in the MDR, and Storage refers to the 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 Oracle RDBMS as its underlying storage mechanism and is based on the models and standards discussed above. The models and standards describe the metadata content and how that content is organized for storage.

4.2 Collection

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.

4.3 Registration

Registration is the process of providing the MDR with its knowledge about the metadata, e.g. name, location, type, etc. The general classes of items which need to be registered are data elements, surveys, products, datasets, and documents. Registration requires several things:

- all the necessary attributes are specified;

- all the necessary links are made (e.g. linking a dataset to all the data elements in its data dictionary);

- classifying the registered item.

Registration tools will have to be designed, probably one for each class of item. The tools will require a template for the user to supply the necessary attributes and make the links to other metadata as needed. Appropriate classification structures will need to be accessible through the tool so each item can be classified.

Useful classification schemes already exist which can be incorporated into registration tools, such as

- TOC;

- Themes as specified in the Cultural and Demographic Data Metadata draft standard of the Federal Geographic Data Committee;

- Thesauri from Statistics Canada and the University of Essex (U.K.).

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

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.

4.4 Retrieval
Retrieval refers to querying metadata in the MDR. Querying will be part of the design of General Purpose Browsers and 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 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.

The TOC view of the SDSM can be used as a check list for categories of metadata. For users wishing to find information about surveys, searching the TOC for the appropriate subject (e.g. questionnaire design) will be useful. Since the TOC Process Model 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.

A prototype metadata browser for the MDR has been built, and browsers are being built for the DADS and FERRET data dissemination systems.

4.4 System Integration
In addition to the tools for collecting and querying metadata, the integration of the MDR with other SIS's needs to be seamless. Two general possibilities for accomplishing this exist. First, the TOC can be used.

A mapping exists between the TOC and the MDR model, and maps can be built from the TOC to the other SIS's by mapping the TOC to 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).

Another solution, probably more effective, is for developers of SIS's to adopt the MDR model for the metadata portion of the SIS. If every SIS at the BOC uses the MDR model, then a distributed metadata repository (each piece based on the same model) will be built. Tools designed to search the metadata in one SIS will be able to search the metadata in all SIS's. A seamless view of the metadata for the entire agency will result. Users who look for BOC data in FERRET will be able to locate data that is only accessible through DADS without having to know which tool to go to first. The actual viewing or downloading of the data will probably require switching tools, but that problem should be minor.

4.5 Metadata Administration
The adoption of the MDR model for storing metadata will require more than supplying information about data elements, surveys, documents, or datasets. Metadata administration is the active management of the information about all the agency's metadata. No function of this type exists at the BOC at this time at the agency level.

The registration process described in section 4.3, and the DER described in section 3.3.2, define generally the information that is required for accurate and complete data administration. The MDR model has expanded the notion of data registration to include metadata.

The registration tools discussed above will handle the entering of metadata into the MDR, but there is a human side to metadata administration which must not be lost in the discussion of the MDR. Some of these 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 MDR a better tool, and facilitate the sharing and understanding of data and metadata among groups within the BOC or with other agencies.

4. Prototype

A series of prototypes is currently under development. The first version is complete. It implemented a subset of the MDR model and contained some information about some data elements and documents. A browser tool was developed using the TOC as a search mechanism for specific types of documents. The browser is a Web based tool that uses a combination of basic HTML, CGI-Perl scripts, and JAVA.

The second prototype is under development now. It is expected to be complete in July. It will implement the complete MDR model, contain substantially more documents, and use an improved version of the browser. Two important functions will be demonstrated: the ability to find metadata across surveys and a tool to register metadata for products. Subsequent prototypes will add more functionality each time.

Usability testing is planned for some of the prototypes. Both the registration tools and the browser will require user feedback to ensure that the tools are useful for users. Unfortunately at this time, the prototypes cannot be released on the Web to the Internet. Much of the metadata in the MDR is not available for the public, and the security functions for the MDR have not been developed to the point where this information is secure.

5. Conclusion

This paper has discussed the work at the BOC to design and build a prototype statistical metadata repository (MDR) using standards developed by international, national, and U. S. Government organizations. Detailed data and metadata models have been built and integrated. The integrated model is the basis for the MDR architecture. It provides a structure for storing the metadata which describes survey designs, processing, analyses, and datasets. The model supports the card catalog metaphor for organizing the BOC metadata.

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.

The first MDR prototype has been built and subsequent ones are planned. Registration and query tools are being developed, and the prototype MDR is being populated with metadata. Increasing interest in using the MDR model for storing metadata for various projects has increased the chance that a seamless distributed metadata repository for the BOC can be developed. Further research, planning, and work will be necessary to bring this plan to reality.

6. 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.

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

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.

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.

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", 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.

* Paper presented at IASSIST/IFDO '97, Odense, Denmark, May 6-9,1997.

Appendix A: Entity Definitions for Business Data Model

Entity Name	Entity Definition	Entity Note
Data Element	A single unit of data that in a certain context is considered indivisible. It cannot be decomposed into more fundamental segments of data that have useful meanings within the scope of the enterprise.	Data is a representation of facts, concepts, or instructions in a form that allows them to be collected, organized, processed and stored in a retrievable form for communication, interpretation, or processing by human or automated means.
Emprise (Project)	An identifiable effort to generate deliverables NOT specific to a single Survey Instance	This appeared in prior models as Project
Emprise_Dataset (Project_Dataset)	A dataset containing either case level data, aggregation of case level data, or statistical manipulations of either.	This appeared in prior models as Project_dataset
Frame	A dataset containing all the cases identified for a Survey Instance based on a Survey's Universe definition	
Methodology	A structured approach to solve a problem	
Product	A finished deliverable of a Project or Survey Instance for external use.	
Program	A group of Surveys related by a common theme. A Program can be made of other Programs	
Purchaser	An external organization or individual who buys Census Bureau products	
Question	A request for one or more related pieces of information from a Case. A Question can contain other Questions	
Questionnaire	An identifiable instrument containing Questions for a particular Survey Instance	

Sample	A dataset containing a subset of a Frame for a particular set of Survey Instances, selected with a specific sampling Technique. For a census, the Sample incorporates the entire Frame.
Supplied_Dataset	A dataset acquired from sources outside the Bureau of the Census. Can be case level or aggregated/transformed data.
Supplier	An external organization which provides data to augment the Census Bureau's efforts
Survey	An investigation about the characteristics of a given Universe
Survey-Dataset	A dataset containing either case level data, aggregation of case level data, or statistical manipulations of either the case level or aggregated survey data, for a single Survey Instance
Survey-Instance	An identifiable activity which uses a System(s) to gather and process a set of Data Items from an identifiable set of cases, for a defined period of time, resulting in one or more specific deliverables
System	An identifiable process, either fully automated or computer assisted, which implements one or more Techniques to produce one or more deliverables. A System can be composed of Systems
Technique	An identifiable algorithm which is used to implement all or part of a Methodology
Universe	The total defined set of interest to one or more Surveys

Appendix B: Entity Definitions for Data Element Registry Model

Entity Name	Entity Definition
Administered Data Registration Component	A generalization for a data element, value domain, data concept, object class or property.
Classified Data Registration Component	A subtype of Administered Data Registration Component, all the data components that require classification.
Conceptual Domain	The set of possible valid values of a data element expressed without representation.
DRC Name Context	An association between an Administered Data Registration Component and a Name Context.
DRC Registration Authority	A registration authority that has registered a particular Data Registration Component.
Data Element	A single unit of data that in a certain context is considered indivisible. It cannot be decomposed into more fundamental segments of data that have useful meanings within the scope of the enterprise.
Data Element Concept	The human perception of a property of an object set, described independently of any particular representation.
Data Registration Classification Scheme	Classification schemes which are used to classify registered data.
Datatype	A category used to classify the collection of letters, digits, and/or symbols to depict values of a data element based upon the operations that may be performed on the data element.
Derivation Type	An entity used to define different types of derivations. Used to normalize the

	derivation type attribute associated for derived Data Elements and Data Element Concepts.
Derived DEC to Derived DE Mapping	An association that tracks a derivation mapping at the conceptual level to a derivation mapping at the data element level if such a mapping were to exist. This is not REQUIRED for all derivation mappings.
Enumerated VD	A list of all permissible values.
Formula	An entity that represents an algorithm to compute values. Formulas involve input quantities (Data Elements) and produce output quantities (Data Elements).
Keyword	An entity that expresses potential search keywords that users of the registry will use to search for and access Data Element Concepts.
Name Context	The system, database, standard document, or other environment in which the logical metadata class functions and the name has meaning.
Non Enumerated VD	A range used for specifying the lower limit and the upper limit of permissible values.
Object Class	A set of concepts, abstractions, or things in the natural world that can be identified with explicit boundaries and meaning and whose properties and behavior all follow the same rules.
Organization	An accredited agency authorized to declare logical metadata classes as registered. (From earlier definition of Registration Authority).
Permissible Values	Allowed values in a Value Domain
Property	A classification of any feature that humans naturally use to

	distinguish one individual object from another. It is any one of the characteristics of an object class that humans use as a label, quantity or description
Registration Authority	The organization authorized to register entries in the Registry.
Representation Class	A classification of value domains based upon the type of representational form.
Synonym Lists	A relationship that captures the fact that two distinct Data Elements have different names but the same meaning (synonym).
Value Domain	A set of Permissible Values, used to represent a Data Element.
Value Meaning	Meaning associated with Permissible Values in an Enumerated Domain.