METADATA STANDARDS AND METADATA REGISTRIES: AN OVERVIEW

Bruce E. Bargmeyer, Environmental Protection Agency, and Daniel W. Gillman, Bureau of Labor Statistics
Daniel W. Gillman, Bureau of Labor Statistics, Washington, DC 20212 gillman_d@bls.gov

ABSTRACT

Much work is being accomplished in the national and international standards communities to reach consensus on standardizing metadata and registries for organizing that metadata. This work has had a large impact on efforts to build metadata systems in the statistical community. Descriptions of several metadata standards and their importance to statistical agencies are provided. Applications of the standards at the Census Bureau, Environmental Protection Agency, Bureau of Labor Statistics, Statistics Canada, and many others are provided as well, with an emphasis on the impact a metadata registry can have in a statistical agency.

Standards and registries based on these standards help promote interoperability between organizations, systems, and people. Registries are vehicles for collecting, managing, comparing, reusing, and disseminating the designs, specifications, procedures, and outputs of systems, e.g., statistical surveys. These concepts are explained in the paper.

Key Words: Data Quality, Data Management

1. INTRODUCTION

Metadata is loosely defined as *data about data*. Though this definition is cute and easy to remember, it is not very precise. Its strength is in recognizing that metadata is data. As such, metadata can be stored and managed in a database, often called a *registry* or repository. However, it is impossible to identify metadata just by looking at it. We don't know when data is metadata or just data. Metadata is data that is used to describe other data, so the usage turns it into metadata. This uncovers the weakness of the definition stated above. We need to invoke a *context*, i.e. a point of reference, to identify what we mean by metadata in a given situation. We need to state precisely which data will be used as metadata for our context.

Metadata management refers to the content, structure, and designs necessary to manage the vocabulary and other metadata that describes statistical data, designs, and processes. This includes the development of metadata models to define the content of metadata within some context, building metadata registries to organize the metadata defined in the model, developing statistical terminologies which define and organize terms into a structure with relationships (e.g., a thesaurus), and identifying the relationships between the terminology structure and other metadata and data.

Much work is being accomplished in the national and international standards communities, especially ANSI (American National Standards Institute) and ISO (International Organization for Standardization). to reach consensus on standardizing metadata and registries. This work has had a large impact on efforts to build metadata systems in the statistical community. Several metadata standards are described, and their importance to statistical agencies is discussed. Applications of the standards at the Census Bureau, Environmental Protection Agency, Bureau of Labor Statistics, Statistics Canada, and many others are described. Emphasis is on the impact a metadata registry can have in a statistical agency.

Standards and registries based on these standards help promote interoperability between organizations, systems, and people. Registries are vehicles for collecting, managing, comparing, reusing, and disseminating the designs, specifications, procedures, and outputs of systems, e.g., statistical surveys. These concepts are explained in this paper.

Metadata helps users understand the meaning and quality of data, and registries and the policies put in place for administering them are used to measure and maintain the quality of the metadata. The connection between good metadata and data quality is described, and an overview of procedures for ensuring metadata quality through metadata administration is discussed.

Many people and organizations that plan to implement standards run into a common problem; there are so many standards it is hard to choose the "right" ones. This paper is an attempt to clarify this situation regarding the management of metadata within the framework of statistical agencies. It is an overview of existing standards that can work together to organize metadata and link them to statistical data and processes.

The paper includes a general description of metadata and metadata registries; a description of metadata management standards; how metadata affects data quality and some measures for quality of metadata itself; and the benefits of implementing metadata standards and registries.

2. STATISTICAL METADATA AND REGISTRIES

2.1 Statistical Metadata

The context we have in mind for metadata in this discussion is statistics. In particular we are interested in the data that are collected and processed through surveys. S*tatistical data*, the data collected and processed through surveys, is called microdata, macrodata, or time series. So, we define *statistical metadata* as the data and documentation that describe statistical data over the lifetime of that data. For the rest of this paper, we will use the term metadata to mean statistical metadata except where noted.

2.2 Metadata Registries

A *metadata registry* is a database used to store, organize, manage, and share metadata. Traditionally, survey groups manage their metadata in their own ways. For example, data dictionaries are created to describe the data elements contained in statistical data sets. There is some coordination of these dictionaries over time but almost no coordination across surveys. A metadata registry is designed to solve this problem by managing metadata from the organization perspective rather than just small program areas. A metadata registry provides for metadata needed to describe objects of interest. It also provides the entities necessary for registration and standardization of those objects (e.g., data elements).

Metadata and metadata registries have two basic purposes (see Sundgren, 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 design and 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.

The efficient and effective use of data is made possible by the organized storage and use of metadata. Data sets become much more useful when useful metadata descriptions are readily available. When metadata are centrally maintained for collections of data sets, users who need to determine which files are appropriate for their work can do so. Many types of requests can be answered through metadata queries. Some examples are:

- Which data sets contain specific information, such as yearly income;
- Which data sets share common information from which links can be made to form larger data sets;
- Locate data sets by broad subjects through pointers to specific items under those subjects;
- Monitor data storage system usage by tracking file sizes;
- Locate surveys with similar or a specific set of characteristics.

3. TERMINOLOGY

A *terminology* is a set of *terms*, which in turn is a word or phrase used to designate a *concept*. A *concept* is a unit of thought (see ISO 704). At the most basic level, a terminology is a listing, without structure, or without reference to a specific language. However, concepts must be *defined* in a natural language, possibly employing other terms. The instantiation of a terminology implies a context. In this case concepts, though not necessarily bound to a particular language, are influenced by social or cultural biases reflected in differences among languages (see ISO 1087-1). Therefore, a terminology may be very specifically related to a subject area, such as retail trade economic surveys.

Terminologies, or classifications, are most useful when a structure is applied to them. The major categories of structure types are thesaurus, taxonomy, and ontology. These are structure types because different kinds of relationships and axioms can be used to build a specific structure for an application. Only the thesaurus structure type has been standardized, but there are draft standards for some types of ontologies. All three structure types are defined below:

- *Thesaurus* - A controlled set of terms covering a specific domain of knowledge formally organized so that the *a priori* relationships between concepts are made explicit;
- *Taxonomy* - A classification according to presumed natural relationships;
- *Ontology* - A formal specification of a conceptualization (i.e., the objects, concepts, and other entities that are assumed to exist in some area of interest and the relationships that hold among them).

Taxonomies are somewhat more restrictive than the other two types. They often contain hierarchical relationships, for example the *North American Industrial Classification System*. However, it is the relationships in a structure that transform a terminology from a simple list to something more meaningful and usable. The breadth and depth of the required relationships help determine the structure type that is needed.

Terminologies help users understand some subject field. The terms and their definitions are the special language that experts in the field use to describe their work and data. Often, though, experts use different terms to mean the same or similar things. In survey statistics, for instance, the terms *frame* and *sample* sometimes mean the same thing. Unless the terms are properly defined, including each of the variants, and useful relationships are established linking the variants and similar terms, confusion will result.

Data elements are the common link across data sets over time and surveys. To understand the content of a data set, one must understand the data elements that make up its data dictionary. Additionally, each survey questionnaire is related to a set of data elements, as are universes, frames, samples, and other design issues.

Terminology is what ties data elements together. Each part of a data element, concept (i.e., definition) and value domain (allowed values), can be described by linking terms to them. The more precisely the terms are defined, the better the terms represent the meaning of the data element. So, users seeking to find data that meets their needs can use terminology to help them find it. A data manager can use terminology to help determine whether data elements mean the same thing, and possibly perform data harmonization. See Figure 1 for the relationship between terminology and other metadata objects.

4. TERMINOLOGY STANDARDS

This section contains descriptions of some standards for terminology. The most general standards apply to thesaurus development, but there are standards for some types of ontologies. The authors are not aware of any standards that describe taxonomies.

4.1 ISO 704

Principles and methods of terminology. This is a standard about how to write standards for terminology. In a sense, it is the most fundamental of the standards we visit in this section. It is divided into three major sections: *concepts*; *definitions*; and *terms*. These are the basic constructs (described in section 5) that are necessary for terminology. Each is described in more detail in the next paragraphs.

An *object* is an observable phenomenon, and a concept is a mental construct serving to classify those objects. Any object may have multiple concepts associated with it, depending on the context or point of view. A *characteristic* is used to differentiate concepts in a terminology, and different types of characteristics are described. The totality of characteristics for a concept is called its *intension*. The totality of objects sharing all the characteristics of a concept is its *extension*. *Relationships* are described in some detail because concepts are always related to other concepts in a terminology. Finally, *systems of concepts* are described. A system is the set of concepts of a given subject field.

Definitions are described in detail. The term is defined, and the purpose of definitions is described. Types of definitions are listed. Within a system of concepts, a definition should be consistent and fix the concept in the proper position within the system. This is called *concordance*. Detailed principles for developing definitions are described (see also Part 4 of ISO/IEC 11179 below). These principles are general and can be used in many situations. Finally, the use of examples is explained.

Terms are also described in detail. First they are defined. The structure and formation of terms is laid out. This includes constituents (or elements) of terms and methods of forming them. Systems of terms are a coherent set corresponding to the system of concepts the terms represent. The correspondence between a concept and term is described. Emphasis is placed on the difficulty of this in natural language. Requirements for the selection and formation of terms are: linguistically correct, accurate, concise, and some other specialized requirements. Finally, abbreviations are discussed.

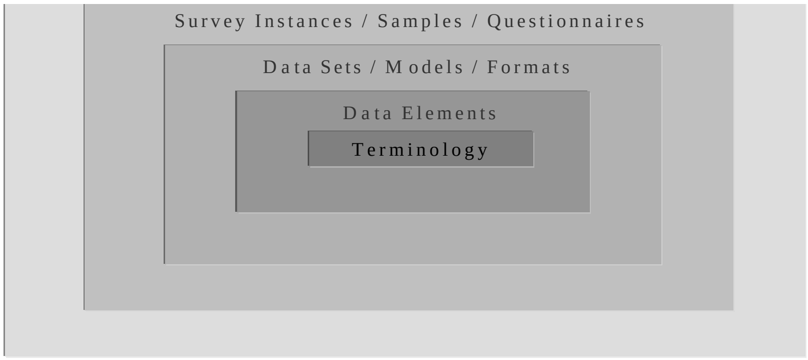
Survey Instances / Samples / Questionnaires
Data Sets / Models / Formats
Data Elements
Terminology

Figure 1: Relationship of Terminology to Metadata

4.2 ISO 860

Terminology Work: Harmonization of Concepts and Terms. This standard specifies a methodology for the harmonization of concepts, definitions, terms, concept systems, and term systems. It is a natural extension of ISO 704.

The standard addresses two types of harmonization: *concept harmonization* and *term harmonization*. Concept harmonization means the reduction or elimination of minor differences between two or more closely related concepts. Concept harmonization is not the transfer of a concept system to another language. It involves the comparison and matching of concepts and concept systems in one or more languages or subject fields.

Term harmonization refers to the designation of a single concept (in different languages) by terms that reflect similar characteristics or similar forms. Term harmonization is possible only when the concepts the terms represent are almost exactly the same. The standard contains a flow chart for the harmonization process and a description of the procedures for performing it.

4.3 ISO 1087-1

Terminology Work: Vocabulary - Part 1: Theory and Application and Terminology Work: Vocabulary - Part 2: Computational Aids in Terminology. These standards establish terminology for doing terminology work. Some of the terms defined in these standards are used throughout this paper. ISO 1087 creates common sets of terms (a special language) for the subject field of terminology application development. Part 1 of the standard identifies, defines, and classifies terms used to develop terminologies for specific applications. The main sections (classes) of the document are: Language and reality; Concepts (includes concepts, characteristics, and relationships); Definitions; Designations (including symbols, names, and terms); Terminology; Aspects of terminology work; Terminological products; and Terminological entries

4.4 Other Standards

There are other standardization efforts for terminology structures and knowledge management. The most important of these are:

- ISO 2788 - Guidelines for the Establishment and Development of Monolingual Thesauri. This standard describes how to build a monolingual thesaurus (i.e., a thesaurus using one language). It uses the theory and structures defined in the standards described above.
- ISO 5964 - Guidelines for the Establishment and Development of Multilingual Thesauri. ISO 5964 sets guidelines for thesauri that have need for more than one language. The content of ISO 2788 is assumed to hold for multilingual thesauri as well as monolingual ones.
- ISO/IEC 14481 - Conceptual Schema Modeling Facility (CSMF).
  This standard defines the constructs to be contained in a modeling facility that can be used to create a formal description of some part of an enterprise. This may include a terminology. A modeling facility is the basis for building a language and a tool to support the activity of creating a formal description. The modeling facility defines the semantics of the language as a set of constructs and how the constructs are related, but not the language syntax.
- NCITS Project 1058 - Knowledge Interchange Format.
- NCITS Project 1059 - Conceptual Graphs

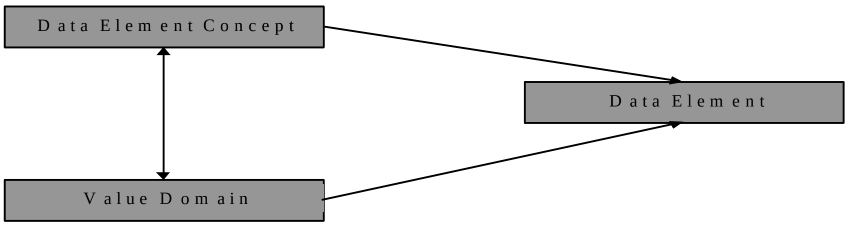

Figure 2: High Level Entity-Relationship Diagram of Metadata Registry

Both of these projects are formal knowledge representation efforts. They and CSMF are part of attempts to standardize ways for a computer to store and use knowledge, described in a formal way (i.e., using first order logic or equivalent schemes). None of these efforts are finished yet, so it is hard to recommend their use. However, they fit into the ontology arena and are part of the larger Artificial Intelligence area of research.

5. ISO/IEC 11179

Specification and Standardization of Data Elements. ISO/IEC 11179 is a description of the metadata and activities needed to manage data elements in a registry. Data elements (or variables) are the fundamental units of data an organization collects, processes, and disseminates. Metadata registries organize information about data elements, provide access to the information, facilitate standardization, identify duplicates, and facilitate data sharing. Data dictionaries are usually associated with single data sets (files or databases), but a metadata registry contains descriptions of the data elements for an entire program or organization.

An important feature of a metadata registry is that data elements are described by a concept (data element concept) and a representation or value domain (set of permissible values). The advantages of this are 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. NAICS 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.

Figure 2 is a high level Entity-Relationship diagram (i.e., a model) for a metadata registry of data elements.

Data elements are described by object class, property, and representation. 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. 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. New object classes are sometimes created 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”.

The *property* is a peculiarity (or characteristic) common to all members of an object class. They are what humans use to distinguish or describe objects. Examples of properties are color, model, sex, age, income, address, price, etc. Again, properties may need to be described using multiple words, depending on the natural language in use.

The *representation* describes how the data are represented (i.e., the combination of a value domain, data type, and, if necessary, a unit of measure or a character set). 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 can also be seen to be composed of two parts: a data element concept and a representation.

Figure 3, below, illustrates the ideas discussed above. It shows that each data element concept is linked to one or more data elements; an object class may be generated from other object classes; a data element concept has one object class and one property associated with it; and a data element has one data element concept and one representation associated with it.

ISO/IEC 11179 - Specification and Standardization of Data Elements - 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. This document is an *International Standard*.
- Part 2 - *Classification of Data Elements* - Describes how to classify data elements. This document is an *International Standard.*
- Part 3 - *Basic Attributes of Data Elements* - Defines the basic set of metadata for describing a data element. This document is an *International Standard*. It is currently being revised.
- 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*.

The revision of ISO/IEC 11179-3 (Part 3) will include a conceptual model for a metadata registry (for data elements). The metamodel, or metadata model for data elements, provides a detailed description of the types of information which belong to a metadata 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. Figure 2 is high level view of the metamodel.

There are two additional important features of the metamodel. It provides for the *registration* of metadata. Registration is a process that administers metadata. It keeps track of who submitted the metadata, who is responsible for it, and the quality of the metadata provided. The ability to measure metadata quality is very important and is an issue which is often overlooked. The other important feature is that the metamodel contains a common metadata entity called *administered_ component*. This entity captures the metadata common to all objects described in the registry. From the perspective of the common metadata, the *administere_ component* entity is like a library card catalog. This makes searching for some objects much easier.

Finally, the revision of ISO/IEC 11179-3 (Part 3) will be more than a data element standard; it will be a metadata registry standard. Many agencies, including the U.S. Census Bureau, are implementing these standards as part of their metadata repository design efforts.

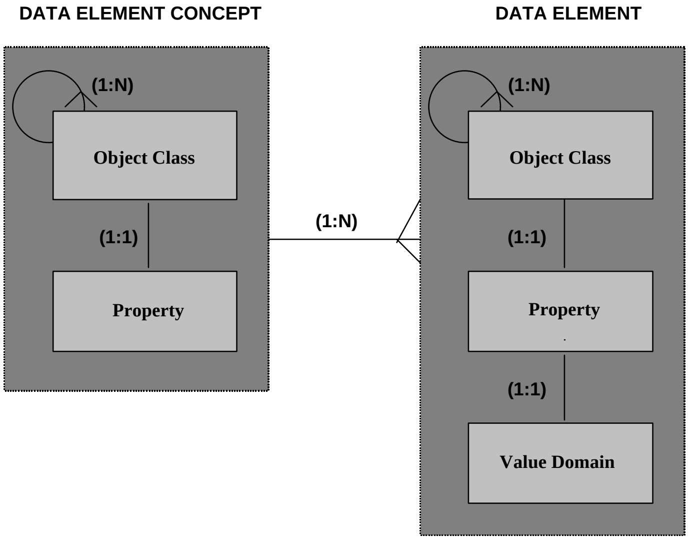

Figure 3: Fundamental Concepts of Data Elements

6. IMPLEMENTATIONS OF ISO/IEC 11179

Many organizations are implementing metadata registries based on ISO/IEC 11179. This section contains descriptions of several of these efforts.

6.1 The Intelligent Transportation System Data Registry Initiative

The Department of Transportation is working on developing 50 to 60 standards for interoperability among the systems that will comprise the nation's Intelligent Transportation System. Five Standards Development Organizations are cooperating in the development of these standards, in addition to identifying a controlled vocabulary and various access paths for information. A major project within this initiative is the Intelligent Transportation System (ITS) Data Registry. It is being designed and developed by IEEE[1] and is based upon the ISO/IEC 11179 standard's concepts. The Transportation Department's explicit use of ISO/IEC 11179 as part of their standards program for the intelligent highway projects has been for the purpose of encouraging and enabling the transportation agencies of the 50 States and Territories to be able to exchange and work with data that has consistent semantics across the various governmental and other organizations involved.

6.2 The Environmental Data Registry

The Environmental Protection Agency (EPA) is developing methods to: a) share environmental data across program systems; b) reduce the burden of reporting regulatory and compliance information; c) improve access to environmental data via the web; and d) integrate environmental data from a wide variety of sources. To accomplish these improvements, the Environmental Information Office is creating the Environmental Data Registry (EDR). The EDR is the Agency's central source of metadata describing environmental data. In support of environmental data standards, the EDR offers well-formed data elements along with their value domains. The EDR is also a vehicle for reusing data elements from other data standards-setting organizations.

The EPA Administrator established the EDR as the agency resource for describing and standardizing environmental data. The EDR supports several strategic goals of EPA, including One Stop reporting, the Reinvention of Environmental Information, and the Public Right to Know. It is used for describing environmental data found inside

[1] Institute of Electronics and Electrical Engineers

and outside of EPA. It is also used by state environmental cleanup program offices to facilitate sharing data among themselves - data that are held only by states and not reported to EPA. The EDR is used to record the results of discussions that rage between program offices about data content and design. It is populated with metadata describing a wide spectrum of environmental data including data in environmental information systems, environmental Electronic Data Interchange (EDI) messages, an environmental data warehouse, environmental regulations, etc. Well-formed data are registered for voluntary use. Mandatory data standards are registered for Agency-wide implementation. The EDR is accessible from the World Wide Web and each month serves up hundreds of thousands of pages. Users download metadata for data elements and groups of data elements. Users also download the entire registry contents.

The registry initiative is engaging European environmental organizations for joint U.S. and European sharing of worldwide environmental data. A major EPA effort is now underway to work on terminology. EPA is building a prototype terminology reference system, as a module of the EDR, that will be compatible with the European environmental terminology system, the General Multilingual Environmental Thesaurus (GEMET). The EDR system provides a direct link to the prototype EPA Terminology Reference System (TRS). The TRS is a new web-based tool for managing lists of terms to be used in data elements, classification systems, ontologies, as well as in text web pages, and other electronic records. Currently, the TRS houses the General European Multilingual Environmental Thesaurus (GEMET).
See: http::/www.epa.gov/edr, http://www.epa.gov/trs and http://www.epa.gov/crs

6.3 Australian National Health Information Knowledgebase

The Australian Knowledgebase is an electronic storage site for Australian health metadata, and includes a powerful query tool. You can use the Knowledgebase to find out what data collections are available on a particular health-related topic or term, and any related official national agreements, definitions, standards and work programs, as well as any linked organizations, institutions, groups, committees or other entities. The Knowledgebase provides direct integrated access to the major elements of health information design in Australia:

- The National Health Information Model;
- The National Health Data Dictionary;
- The National Health Information Work Program; and
- The National Health Information Agreement.

The Knowledgebase does not, as a rule, provide access to actual data through its searching and querying tools, but it is a planned future development.
See: http://www.aihw.gov.au/services/health/nhik.html

6.4 The United States Health Information Knowledgebase

The Department of Defense - Health Affairs (HA) in collaboration with the Health Care Financing Administration (HCFA) is developing the United States Health Information Knowledgebase (USHIK) Data Registry Project. The project goal is to build, populate, demonstrate, and make available for general use a data registry to assist in cataloging and harmonizing data elements across multi-organizations. The requirements team includes representatives from the Department of Veteran Affairs, the Health Level Seven (HL7) standards committee, the Health Care Financing Administration, and the Department of Defense Health Affairs office. The implementation builds on Environmental Protection Agency and Australian Institute of Health and Welfare implementations and utilizes DoD - Health Affairs' Health Information Resource Service (HIRS) to develop and implement a data registry. The project utilizes selected Health Insurance Portability and Accountability Act (HIPAA) data elements for demonstration. The data elements are those used in standards by the X12 (EDI[2]) standards committee, the HL7 standards committee, the National Council of Prescription Drug Program (NCPDP), and the National Committee on Vital and Health Statistics (NCVHS).

An EPA, HCFA, and DoD-HA joint effort also is an initial model for interagency agreements and working arrangements can be made between agencies for synergistic development of metadata and data registry technology.
See: http://hmrha.hirs.osd.mil/registry/

6.5 The Census Bureau Corporate Metadata Repository

The Census Bureau is building a unified framework for statistical metadata. The focus of the work is to integrate ISO 11179 and survey metadata, using the metadata to enhance business applications. A production corporate

[2] Electronic Data Interchange

metadata registry (CMR) is under development based on an extended model including ISO 11179 and a business data model. The goal is put metadata to work to guide survey design, processing, analysis, and dissemination.

The development process features pilot projects that use the CMR from different subject areas within the agency. Each of these projects focuses on a different aspect of the CMR and the information potential it provides. As more projects use the CMR, its visibility and usefulness within the agency will increase.

Current project applications include the American Fact Finder - Data Access and Dissemination System. This project is a large effort to disseminate Decennial Census, Economic Censuses, and American Community Survey data via the Internet. Other projects are integrating the CMR with electronic questionnaire design tools, batch and interactive CMR update and administration tools, and interfaces with statistical packages such as SAS.

6.6 Statistics Canada Integrated MetaDataBase

Statistics Canada is building a metadata registry, called the Integrated MetaDataBase, based on the same conceptual model for statistical metadata developed at the Census Bureau. This effort is still in the design and initial implementation stages. It will integrate all the surveys the agency conducts, contain many standardized and harmonized data elements, and link statistical data to the survey process.

6.7 OASIS and XML.org XML Registry and Repository Work

OASIS, the Organization for the Advancement of Structured Information Standards is developing a specification for distributed and interoperable registries and repositories for SGML[3] and XML[4] DTDs (Document Type Definitions) and schemas. XML.org, an initiative within OASIS, is implementing one such registry and repository. The basis for much of the specification is ISO/IEC 11179.
See: http://www.oasis-open.org/html/rrpublic.htm

7. REGISTRATION AND QUALITY

The quality of data is enhanced when the proper metadata is available for that data. Data is more understandable and useable in this case. Also, data quality statements themselves are metadata. So, metadata describing sampling errors, non-sampling errors, estimations, questionnaire design and use, and other quality measures all need to be included in a well-designed statistical metadata registry. However, this does not say anything about the quality of the metadata itself.

Registration is the process of managing metadata content and *quality*. It 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 statuses) 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".

Semantic content addresses the meaning of an item described by metadata. Usually the name and definition are the extent of this, but for statistical surveys, much more relevant information is necessary to describe an object. A data element has a definition, but additional information that is necessary to really understand it is: the value domain; the question that is the source of the data; the universe for the question; the skip pattern in the questionnaire that brought the interviewer to the question; interviewer instructions about how to ask the question; sample design for the survey; standard error estimates for the data; etc. The model driven approach to the CMR is a start to understanding this.

Once the semantic content is really known, then the work to harmonize some data across surveys and agencies can begin. Harmonization can occur at many levels, e.g., data, data set, and survey.

[3] Standard Generalized Mark-up Language

[4] eXtensible Mark-up Language

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. As such, metadata 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.

Research will focus on ways of measuring quality using these and other criteria.

8. BENEFITS

The benefits of implementing an ISO/IEC 11179 metadata registry are several:

- Central management of metadata describing data and other objects throughout the agency;
- Increased chances of sharing data and metadata with other agencies that are also compliant with the standard;
- Improved understandability of data and survey processes for users;
- Single point of reference for data harmonization;
- Central reference for survey re-engineering and re-design.

The structure of an ISO/IEC 11179 metadata registry has many points at which terminology will aid in searching and understanding the objects described. As described above, terminology is developed to aid the user, designer, and analyst. The *meaning* of an object is clarified through the set of all the terms linked to it. Although a definition is important for understanding an object, it often does not convey the full context in which the definition is made. A good example is a question in a questionnaire. The question wording itself serves as its definition, but the universe for which the question is asked or about is usually not specified. That context is inferred from the flow and answers to previous questions. However, appropriate terms associated with a question can convey some of this necessary information without resorting to following a complicated questionnaire down to the question under consideration.

In conclusion, many organizations are implementing ISO/IEC 11179 metadata registries, including the U.S. Census Bureau, U.S. Environmental Protection Agency, U.S. Health Care Financing Administration, Australian Institute for Health and Welfare, and others. The international standards committee **ISO/IEC JTC1/SC32/WG2** (Metadata) is responsible for developing and maintaining this and related standards. Participation by national statistical offices through the appropriate national standards organization will make this effort much stronger and provide a means to interoperability across national boundaries for statistics.

9. REFERENCES

- Gillman, D. W., Appel, M. V., and Highsmith, S. N. Jr. (1998), “Building a Statistical Metadata Repository”, Presented at METIS Workshop, Geneva, Switzerland, February, 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.
- ISO 704, Principles and Methods of Terminology, 1987, International standard.
- ISO 860, Terminology Work - Harmonization of Concepts and Terms, 1996, International standard.
- ISO 1087-1, Terminology Work - Vocabulary - Theory and Application, 1995, International standard.
- 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.