Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2005 : Methodological Challenges for Future Information needs

2005

Statistics Canada **Statistique Canada**

Canada

Proceedings of Statistics Canada Symposium 2005
Methodological Challenges for Future Information Needs

COMMON METADATA CONSTRUCTS FOR STATISTICAL DATA

Daniel W. Gillman[1,2]

ABSTRACT

Regardless of the specifics of any given metadata scheme, there are common metadata constructs used to describe statistical data. This paper will give an overview of the different approaches taken to achieve the common goal of providing consistent information that supports the interpretation of statistical data and draw parallels and similarities between them. It will then present how the ISO/IEC 11179 standard addresses the issue.

KEY WORDS: DDI; ISO/IEC 11179; Neuchâtel; SDMX; Terminology Theory

1. INTRODUCTION

For over 15 years, metadata management has been an important concern in statistical offices around the world. Statistical metadata, metadata for statistical data and processes, is used to enhance users' search and understanding of statistical data, improve survey processing within each office, and facilitate statistical data harmonization, among many others. As a result, the area is a fertile ground for research and development. Many offices are using metadata driven systems to automate parts of the survey process (Johanis, 2000; Oakley, 2004)

Several things need to be understood and developed before a metadata driven system can be built. Foremost, an understanding of what constitutes metadata for the problem at hand. Metadata is not an absolute concept. Data are not metadata because of some inherent properties, they are metadata by use. So, metadata is a relative idea. Data become metadata when they are put into a descriptive relationship with something else (Gillman, 2005; Farance and Gillman, 2006).

Once the required metadata elements are understood, a model can be built. This is a data model of the metadata to be used. The model is a framework for how the metadata will be organized in a database, and the structure is often optimized in some way to enhance the uses of the database (Date, 2003). The common constructs among models and their attributes are the focus of the discussion in this paper, as metadata constructs are components of models.

Most situations require some amount of modeling work. It is rare that a model in use in one statistical office will work without modification in another. A reasonable question when designing a model is "Has anyone else thought about this problem, and is there a solution I can borrow that will work for my situation?" Already existing models may not work at all, may work for some purposes but not others, or may work completely. For the models that fit partially, they can be made to work if they can be modified. This is usually the case.

Where does one look for appropriate models? There are 4 possible answers: other statistical offices, commercial software, published papers or books, and standards. Other statistical offices are a great source for metadata models, as several good metadata models have been developed there (Johanis, 2000). Commercial software usually does not have appropriate metadata models, as the needs of statistical offices are too specialized, and commercializing specialized products does not pay off. Metadata models in books and papers are too high level, so not so useful for building systems. However, they are useful for conveying a conceptual framework, which is shared. Finally, standards are a good source for metadata models, because they contain much detail and are based on consensus among a wide group. Standards are often built by a community of practice, people in similar businesses or

[1] Daniel W. Gillman, Bureau of Labor Statistics, Office of Survey Methods Research, Room 1950, 2 Massachusetts Ave., NE, Washington, DC, 20212, USA, Gillman_D@BLS.Gov
[2] The opinions in this paper are due solely to the author and do not necessarily reflect the policies of the Bureau of Labor Statistics.

otherwise having like concerns. This leads to the development of standards that appeal to specialized groups, e.g., the Data Documentation Initiative (ICPSR, n.d.).

Standards and other statistical offices seem to be the best sources for finding appropriate metadata models, and we will analyze four metadata schemes, which arose from each of those sources. Part of the analysis will include a discussion of common constructs. Regardless of the specifics of any given scheme, there are common metadata constructs used to describe statistical data. This paper will give an overview of the different approaches taken to achieve the common goal of providing consistent information that supports the interpretation of statistical data and draw parallels and similarities between them. It will then present how the ISO/IEC 11179 standard addresses the issue.

The paper is organized into substantive sections. We begin with a section on the theory of terminology. This provides a framework in which to conduct the analysis. Next, a discussion of statistical data based on terminology theory is provided. Then, there is a description of the most important constructs for each of four schemes: Data Documentation Initiative (DDI); ISO/IEC 11179; Neuchâtel Variables and Classification models; and Statistical Data and Metadata Exchange (SDMX). Finally, the common constructs, a comparison between the models, and a framework for using the models together in a statistical office are provided.

2. THEORY OF TERMINOLOGY

2.1 Basic Definitions

To begin, we describe some useful constructs from the theory of terminology. These come from several sources (Sager, 1990; ISO, 1999; ISO, 2000). The constructs and their definitions follow below:

- *property* - observation, used to describe or distinguish an object (e.g., "Dan has blue-gray eyes" means "blue-gray eyes" is the property of Dan. It is abstracted to a characteristic, color of eyes, of people - see characteristic.)
- *object* - something conceivable or perceivable
- *characteristic* - abstraction of a property of a set of objects
- *concept* - unit of knowledge created by a unique combination of characteristics
- *intension* - sum of characteristics that constitute a concept
- *extension* - set of objects to which a concept refers
- *definition* - expression of a concept through natural language, which specifies a unique intension and extension
- *concept system* - set of concepts structured according to the relations among them
- *designation* - representation of a concept by a sign, which denotes it
- *general concept* - concept with two or more objects that correspond to it (e.g., planet, tower)
- *individual concept* - concept with one object that corresponds to it (e.g., Saturn, Eiffel Tower)

Designations come in three types: A term is a verbal designation of a general concept; an appellation is a verbal designation of an individual concept; and a symbol is any other designation.

The ancient Greek philosophers began the study of terminology and concept formation in language (Wedberg, 1982), and they discovered a useful relationship between designation, concept, object, and definition, that is illustrated in Figure 1 (CEN, 1995). This diagram, minus the definition part, is often referred to as Ogden's Triangle (Ogden and Richard, 1989).

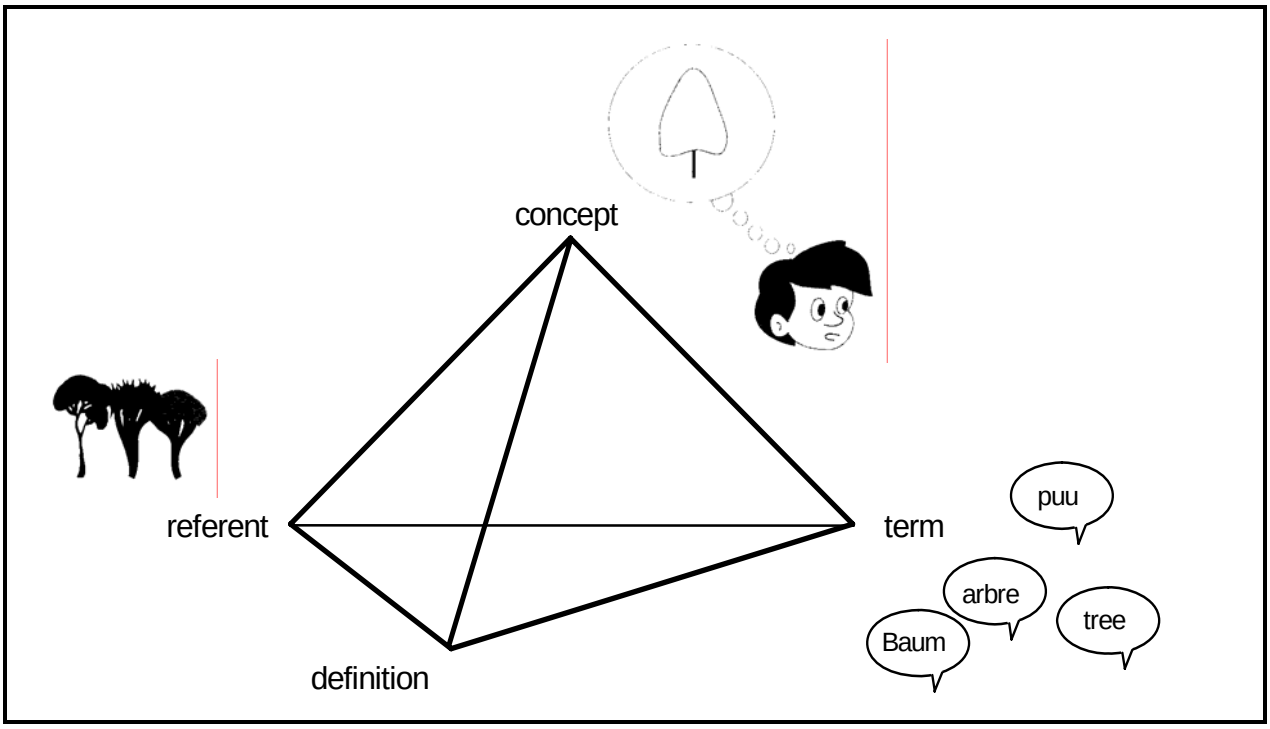

Figure 1. Relationships between referents (objects), concepts, terms (more generally designations), and definitions.

An important observation is that concepts are human constructions (Lakoff, 2002). No matter how well we define a concept, a complete description is often impossible. Identifying the relevant characteristics is culturally dependent. So, some objects in the extension of a concept, called prototypes, fit the characteristics better than others (Lakoff, 2002). For example, an eagle fits more of the characteristics of a bird than a penguin does.

2.2 Data

Another important observation is that a datum is a designation (Farance and Gillman, 2006). A value obtained through the interview process or recorded in a database (actually an instance of a sign - see section 2.1) denotes some concept, therefore is a designation, most likely a symbol.

Statisticians view a datum as a designation of a class in a partition of a population of objects, where the partition[3] is defined for some characteristic of the population (Froeschl, Grossmann, & Del Vecchio, 2003). Here, in a way that departs from traditional statistics theory a little, the population is either a general or individual concept, and the objects are the extension of that concept. Two examples illustrate the ideas:

[3] A partition is a non-empty set of mutually exclusive and exhaustive subsets of some other set. The number of subsets is not necessarily finite.

Example 1: Population as general concept

Population: The set of adults age 16 and older in Canada in 2006 Characteristic: Sex Partition: {Male, Female} Designations: 0 for Male 1 for Female

Example 2: Population as individual concept

Population: The set of the set of adults age 16 and older in Canada in 2006 Characteristic: Proportion of females Partition: {x | 0 <= x <= 1} Designations: Real numbers between 0 and 1, with precision to 3 decimal places

In addition, there are two senses in which the term population is used. First of all, a population is a concept represented by a definition or description. Also, it is the extension of the concept, the set of objects about which we collect or observe data.

As a concept, a population is either individual or general. In the case of microdata, populations are general concepts - see Example 1. However, aggregate data, or macrodata, requires a population with one object. In Example 2, the characteristic "proportion of females" applies to the set containing the set of adults age 16 or older in Canada in 2006, not to each individual. The set consisting of "the set of adults age 16 or older in Canada in 2006" has one element. But, it is this aggregate that has the characteristic "proportion of females", not the individual people. "Proportion of females" is not a characteristic of people, "sex" is. Likewise "sex" is not a characteristic of the aggregate, "proportion of females" is.

When one conjures a particular object in the mind, the conception of that object is an individual concept. This means every object has an individual concept associated with it. Data associated with a particular object is descriptive of that object and this means the data are metadata. Data are only metadata when they are used to describe some object. However, metadata are similar to aggregate data, their populations are individual concepts. This implies all data are metadata at the point of collection!

The framework just described is a basis for describing all statistical data. In the sections that follow, we will describe how well each metadata scheme follows this framework.

2.3 Implementation

The ISO/IEC 11179 standard implements the terminological theory of data in a very straightforward way. For each of the constructs in the theory, there is one in the standard (ISO, 2005)

Without going into details about the model described in the standard and defining all the terms there, we will list the mapping between the terms in the standard and the terms from the terminological theory. For more details, see ISO/IEC 11179-1: *Framework* (ISO, 2005).

Here is the list, with the ISO/IEC 11179 term listed first:

- object class — population
- property[4] — characteristic
- partition — conceptual domain
- categories — value meanings
- designations (values) — permissible values

[4] This was a most unfortunate choice. The term will be changed to characteristic in the next (3rd) edition.

There are many more details between the terminological theory and the standard that can be mapped. However, this would require a much deeper discussion of each, and the paper is meant be an overview.

3. METADATA SCHEMES

In this section, four metadata schemes are described, and comparisons made between them.

3.1 DDI

The Data Documentation Initiative (DDI) is an international project to establish an XML-based metadata standard for the content, presentation, transport, and preservation of documentation for datasets in the social sciences. Social scientists need to record and communicate all the important characteristics of the empirical data for which they are responsible in a straightforward way. The DDI endeavors to do this.

The DDI metadata specification originated in the Inter-university Consortium for Political and Social Research and is now the project of an alliance (http://www.icpsr.umich.edu/DDI/org/index.html) of about 25 institutions in North America and Europe. It is based on the idea of the electronic "codebook," retaining its capabilities, but growing the possibilities by improving the rigor and expanding the scope. The DDI is in use by many social science data archives and statistical offices around the world.

The DDI is represented as an XML DTD and an XML-Schema (W3C, 2004). The DDI-DTD is divided into 5 main chapters:

- Document Description - description of the XML document itself
- Study Description - description of the study behind the data
- File Description - physical layout of the described data set(s)
- Data Description - conceptual description of the data
- Other Material - descriptions of related data and documents

The main focuses of the DDI are the study and the data set from the perspective of social science statistics. The study is the only required chapter, and it represents a high level description of the data. This means that archives can maintain individual descriptions of each data set they manage. However, this also means that some of the metadata for a series of data sets from the same survey or program must be repeated.

The DDI has a rich set of elements devoted to the needs of statisticians and other users of statistical data. It is the only one of the four schemes that is specifically engineered for describing statistical surveys. The Neuchâtel Group, described below, also produces standards directly related to statistical offices, but they are much more specialized.

Under the variable description section of the data description chapter, there are some elements for capturing concepts, but there is little in the way of concept management. Part of this is due to the design. XML is hierarchical, and it is hard to model complex relationship structures. Revisions of the DDI are expected to address some of this.

3.2 ISO/IEC 11179

The ISO/IEC 11179 - *Metadata registries* - standard is a metadata specification devoted to data semantics. It also contains a model and an overview of a procedure for registration, hence the "registries" in the name. However, the main focus is semantics.

The standard was written in six parts:

- Part 1 - Framework -- an overview of the standard and the methodology behind data semantics
- Part 2 - Classification -- presentation of a model for managing a classification scheme, especially as it relates data elements (variables) to each other

- Part 3 - Metamodel and basic attributes -- presentation of the full model for data semantics, classification, and registration
- Part 4 - Formulation of data definitions -- principles for writing good data definitions
- Part 5 - Principles for naming and identification -- provides a naming convention for each of the principal parts of data semantics
- Part 6 - Registration -- procedures for registration

The last published version of the standard is the 2nd edition, completed in 2005. All the latest published parts of ISO/IEC 11179 are freely available on the web site of the Information Technology Task Force (ITTF) under ISO[5] and IEC[6] (http://isotc.iso.org/livelink/livelink/fetch/2000/2489/Ittf_Home/ITTF.htm).

The basic unit in ISO/IEC 11179 is the data element (variable). The model specified in the standard shows how one should describe a data element. It is concept based and follows the general framework of the terminological theory of data described above.

However, the standard does not address statistical data *per se*. It contains a general description of data, and does not go any further than that. Even the idea of a data set is not described in the standard.

3.3 Neuchâtel Group

The Neuchâtel Group is currently composed of representatives from Statistics Netherlands, Statistics Norway, Statistics Sweden, Swiss Federal Statistical Office, and the US Bureau of Labor Statistics. In addition, a small German software company, run-Software AG, is part of the group, and they build software to implement the specifications.

An earlier version of the group developed a model for managing classification systems used in statistical offices. This is called the Neuchâtel Group Classification model. It is in use by many statistical offices in Europe, North America, Australia, and New Zealand.

The Neuchâtel Group Variable Model is still under development. Obviously, the main construct described is the variable or data element. Many of the same components of data contained in ISO/IEC 11179 are contained in the Neuchâtel Group Variable Model. The first version of the specification is due to come out later this year. It is very much based on ISO/IEC 11179, but it has some major differences, too. First, it is developed using statistical terms and contains detail for statistical data that ISO/IEC 11179 lacks. Also, it describes databases, files, and formats in addition to basic variables.

The Variables Model does not contain the capability for registration, and it does not contain all the flexibility that ISO/IEC 11179 has. However, for the description of a single data element, they are very similar.

Since the specification is still in draft, the group does not want the document distributed at the time of writing this paper, so no reference to it is given. Also, because of limited resources, the group does not have a web site. This is also under development.

3.4 Statistical Data and Metadata Exchange (SDMX)

The SDMX project is jointly sponsored by seven international statistical and financial organizations: World Bank, Bank of International Settlements (BIS), International Monetary Fund (IMF), Organization of Economic Cooperation and Development (OECD), Eurostat, European Central Bank (ECB), and UN Statistics Division. Each organization has the need to describe, share, and transfer statistics and their metadata.

The project was begun in 2001, and now the 2nd version of the work is available at the project web site: http://www.sdmx.org. Also, the work is being developed as an international standard, ISO 17369.

---

[5] International Organization for Standardization

[6] International Electrotechnical Commission

The SDMX model is very sophisticated and general. Sufficient analysis of it is not complete, but a major concern is that it was not developed specifically for statistics. There are not enough attributes to situate the model directly in the realm of statistics. As a result, the generality runs the risk of satisfying the needs of many kinds of businesses. That may be a strength, too, as the developers can use the model to transfer many kinds of data and metadata. However, then SDMX may be a misnomer.

The main construct in SDMX is the *key family* and related classes. Many of the components of data are described here. Also, the authors claim that everything in ISO/IEC 11179 may be mapped to SDMX model. This has not been checked in a rigorous way, but there is a strong relationship between the standards.

Registration is also a part of SDMX. The developers followed the design of the ebXML registry specification defined by the Organization for the Advancement of Structured Information Standards (OASIS, 2002). This registry specification was generalized from the ISO/IEC 11179 registration idea and is conceptually similar.

3.5 Comparison

The main construct in common between these models is the statistical variable. Whereas the Neuchâtel Variables Model focuses on microdata and the ISO/IEC 11179 model describes any data, the SDMX model focuses on macrodata and time series, although in the cases of SDMX and Neuchâtel, other kinds of data may be described, too. The DDI describes both microdata and macrodata, but the details of the description are lacking somewhat.

Since ISO/IEC 11179 contains the most general description of data, then it is the central model of the four. As we showed in section 2.3, the ISO/IEC 11179 standard implements the terminological description of data. Based on the discussion in section 2.2, then the ISO/IEC 11179 model is fundamental. This does not mean, especially, that the SDMX and Neuchâtel models do not provide the same functionality. However, each of these is devoted to a specific subject matter domain - statistics. They are designed to do other things for statisticians besides describe statistical data.

How does one map between these standards? A good way is to think of ISO/IEC 11179 as the hub of a wheel and the other specifications as spokes on that wheel. To map from one to another, first map from one to the center and then back out to the other. Given that we are only considering four models, then it is easy to create pair wise maps (There are six.). However, spoke wise maps only require three.

A statistical office can take all four of these specifications to create a complete metadata specification. ISO/IEC 11179 describes registration and data semantics. The Neuchâtel Group Variable Model describes data semantics, data sets, cubes, and databases. The DDI describes data sets, cubes, and statistics. SDMX describes registration and data/metadata transfer. The combination provides a powerful framework.

REFERENCES

CEN. (1995), *Medical Informatics - Categorical Structures of Systems of Concepts*. Draft. Brussels: European Committee for Standardization.

Date, C. (2003), *An Introduction to Database Systems (8th ed)*. Addison Wesly.

Farance, F. and Gillman, D. (2006), "The Nature of Data", Manuscript submitted for publication.

Froeschl, K., Grossmann, W. and Del Vecchio, V. (2003), *The Concept of Statistical Metadata*. Deliverable #5 for MetaNet Project. Retrieved July 2004 from http://www.epros.ed.ac.uk/metanet/deliverables/D5/IST-1999-29093_D5.doc.

Gillman, D. (2005), "Data Semantics", In D. Schwartz (ed.) Encyclopedia of Knowledge Management. Hershey, PA: Idea Group.

ICPSR (Inter-University Consortium for Political and Social Research). (n.d.), *Data Documentation Initiative*. Retrieved July 2004 from http://www.icpsr.umich.edu/ddi.

ISO. (1999), *ISO 704: Principles and methods of terminology*. Geneva: International Organization for Standardization.

ISO. (2000), *ISO 1087-1: Terminology – Part 1: Vocabulary*. Geneva: International Organization for Standardization.

ISO. (2005), *ISO/IEC 11179 - Metadata registries (All Parts)*. Geneva: International Organization for Standardization and International Electrotechnical Commission.

Johanis, P. (2000, November), *Statistics Canada's Integrated Metadatabase: Our Experience To Date*. Invited Paper #3 presented at the UNECE Workshop on Statistical Metadata. Washington, DC.

Lakoff, G. (2002), *Women, Fire, and Dangerous Things* (Reprint edition). University of Chicago Press.

Oakley, G. (2004, February), *Using ISO/IEC 11179 to help with metadata management problems*. Invited Paper #9 presented at the Joint UNECE, Eurostat, OECD Workshop on Statistical Metadata,. Geneva.

OASIS. (2002), *OASIS/ebXML Registry Information Model, v2.0*. Organization for the Advancement of Structured Information Standards. Retrieved January 2006 from http://www.oasis-open.org.

Ogden, C. and Richard, I. (1989), *The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism*. Harcourt.

Sager, J. (1990), A *Practical Course in Terminology Processing*. John Benjamins.

Wedberg, A. (1982), *A History of Philosophy - Vol 1: Antiquity and the Middle Ages*. Oxford: Clarendon Press

W3C. (2004), *Extensible Markup Language*. XML 1.1 reference specification. Retrieved July 2004 from http://www.w3c.org.