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

DOCUMENTING DATA ELEMENTS IN STATISTICAL AGENCIES

Paul Johanis[1]

ABSTRACT

The explanatory information accompanying statistical data is called metadata, and its presence is essential for the correct understanding and interpretation of the data. Over the years, there has been an ongoing search for a standard way of structuring and presenting this type of explanatory information. This paper will report on the experience of Statistics Canada in the conceptualization, naming and organization of variables on which data are produced.

1. INTRODUCTION

The main function of national statistical agencies is to produce and disseminate statistical data on the economic and social conditions in their country. Statistical data take the form of numbers of various types, in data files, statistical tables or in texts such as news releases and articles. These numbers on their own cannot be understood. This explanatory information is called metadata, and its presence is essential for the correct understanding and interpretation of statistical data.

At the most fundamental level, this explanatory information must cover at least the description of the data. A standard that is useful, and that is being used by Statistics Canada, to structure and present this type of metadata is IS0/IEC 11179, *Information Technology - Specification and Standardization of Data Elements.*[2] In statistical terminology, data elements are commonly referred to as variables. This standard therefore provides a guideline for structuring and presenting basic descriptive information about variables. The very process of creating descriptive information according to this standard, however, also has the effect of bringing about more consistency and rigour in the conceptualization, naming and organization of variables for which data are produced. This paper will report on the experience of Statistics Canada in this regard.

2. THE STANDARD

ISO/IEC 11179 is a standard for creating and maintaining data element registries. It is organized in six parts. Part 3 provides the relevant framework for documenting data elements. Statistics Canada has to a large extent implemented the 2003 version of this part of the standard (second edition) in its metadata registry.

The fundamental aspects of this part of the standard are illustrated in the following chart. Data elements are expressions of data element concepts and are represented through value domains. Roughly translated in statistical terminology, this is equivalent to saying that variables are expressions of their underlying concept and are represented through classifications.

---

[1] Paul Johanis, Statistics Canada, Canada, K1A0T6

[2] ISO/IEC 1994, 1995, 1997, 2000, 2001, 2002. NOTE ISO/IEC 11179 is currently being revised under the general title *Information technology — Metadata registries (MDR)*.

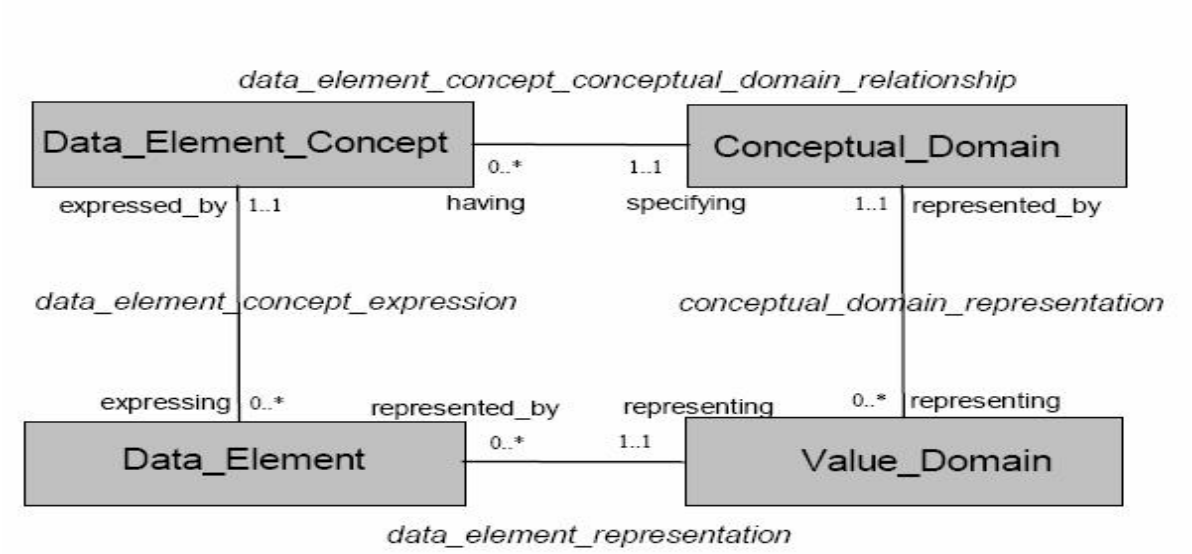

Figure 3 — High-level metamodel

Each of the components of this high level metamodel is further decomposed in the standard. The data element concept is composed of an object class, which in statistical parlance refers to the unit of observation, and a property. The data element is composed of a data element concept and a representation class. Value domains can be non-enumerated, that is they can assume any value in a continuous range, or enumerated, in which case they can assume a given set of permissible values and value meanings. These decompositions will be presented in detail in the context of Statistics Canada's implementation of the standard.

3. APPLICATION IN STATISTICS CANADA'S METADATA REGISTRY

Statistics Canada has implemented a corporate metadata registry, which is known internally as the Integrated Metadatabase (IMDB). It contains information on all of the agency's 400+ active statistical programs. The type of information provided covers the sources and methods used to produce the data published from these programs, indicators of the quality of the data as well as the names and definitions of the variables, and related classifications, published through the agency's online database, known as CANSIM.[3]

For each of these variables, metadata fully specified according to the standard has been created. Therefore, in each case, the object class, the property and the representation class have been specified, named and defined, as have their value domains. Each of these components can stand on its own and is therefore reusable in the construction of other data elements. The general strategy was therefore to be very economical in the creation of these constituent parts and to use combinations and permutations of these elementary components to represent the diversity of variables for which data are published by the Agency.

The process followed to create this metadata started with the analysis of each array of statistical data presented to users of CANSIM on the Statistics Canada website. There are 1500+ such arrays in CANSIM. For each table, the process involved first identifying the unit of observation for the variables contained in the table. This led to the specification, naming and defining of the object class. Having defined the object class, the next step was to identify

[3] The full set of variable level metadata is only viewable within Statistics Canada at this time. Starting July 6, 2005, reviewed and validated variable level metadata is gradually being added to Statistics Canada's externally viewable website for all published variables in CANSIM.

the properties of this unit that were being measured, which were then named and defined. Finally, the type of representation was identified (a count, an index number, a value, a type, etc.). These basic elements provided the necessary building blocks to name and define data elements, in the form *Representation class* of *Property* of *Object Class*. This gives variable names such as Type of Industry of Establishment, or Category of Age of Person, or Index of Producer Price of Product.

For the variables that are represented by values in a classification, the appropriate value domains were then identified, named and defined. Value domains in the standard are equivalent to what would be considered all classes in a given level of a statistical classification. These are also standalone and reusable in the context of many data elements.

All of the building blocks, and the links between them in the context of given statistical tables, were stored in the IMDB, from which it is therefore possible to produce, dynamically and on request, the complete definition of every variable, according to the specifications of the standard. Using the variable Type of Expenses of Business Location as an example, the end result is presented in this format:

Type of Expenses of Business Location

'Expenses' refer to decreases in economic benefits or service potential, during the reporting period, in the form of outflows or consumption of assets or incurrence of liabilities that result in decreases in equity, other than those relating to distributions to owners.

'Business Location' refers to a statistical unit defined as a producing unit at a single geographical location from which economic activity is conducted and for which, at a minimum, employment data are available.

'Type' refers to the reporting of 'Expenses of Business Location' using the following classification(s):

- Expense Categories, Annual Survey of Manufactures (ASM)

As the analysis progressed, many situations arose where conventions or consistent approaches needed to be developed and applied to deal with the multifarious ways used by data producers to present statistical data. This paper will report on the practical choices that were made in applying the standard, including the conventions adopted to deal with some of the more common situations encountered.

4. STATISTICAL UNIT= OBJECT CLASS

Statistical units are defined in Statistics Canada's Policy on Standards as: "The unit of observation or measurement for which data are collected or derived".[4] With reference to the ISO standard, this makes it a statistically relevant type of object class, defined in the standard as: "a set of ideas, abstractions, or things in the real world that can be identified with explicit boundaries and meaning whose properties and behavior follow the same rules."[5] In applying the standard, it was decided to try to limit the number of object classes identified by using wherever possible fundamental statistical units. Fundamental statistical units are defined as those that are not types of any other statistical unit and cannot be derived as a grouping of any other statistical unit.[6] The following types of fundamental statistical units were identified:

- Agents: Entities that act and whose actions are reported on by Statistics Canada. In social statistics, "Person" can be considered as an agent.
- Events: Actions of (or by) agents as reported by Statistics Canada. Events are discrete in time (occur in a time period) and finite (can be counted). In social statistics "Birth" can be considered an event.

[4] Statistics Canada, Policy on Standards, http://www.statcan.ca/english/about/policy/standards.htm

[5]

[6] See Mechanda, K., Johanis, P., and Webber M. (2003) *Conceptual Model for the Definitional Metadata of a Statistical Agency*, Paper for Open Forum 2003 on Metadata Registries, Santa Fe, New Mexico.

- Items: Things that are generally either produced or managed by agents. In economic statistics, "Product" can be considered as an item.

Fundamental statistical units were identified as a means of keeping the number of object classes to a minimum. However, some statistical units that can be derived in some way from the fundamental statistical units are so commonly used that they were identified as separate object classes. Common derivations of the fundamental statistical unit include:

- Subsets of fundamental statistical units based on an inherent characteristic. An example of this is "Person age 15 and over" used as an object class. This is a subset of the "Person" object class based on the Age property. In this way, data elements such as "Type of Occupation of Person age 15 and over" can be defined.
- Subsets based on roles that the statistical units may assume. Examples of this are "Student", "Mother" and "Employee", subsets of the "Person" object class based on various role properties. In this way, data elements such as Category of Major field of study of Student can be defined. Subsets based on roles differ from those based on inherent characteristics in that the same statistical unit can take on more than one role at the same time. It can both assume and discontinue a role over time.
- Supersets of fundamental statistical units. For example, "Family" is a group of persons according to certain grouping rules.

Starting with fundamental statistical units and only identifying additional statistical units defined according to certain properties when these were commonly encountered has led to the identification of a fairly limited set of about 80 object classes.

Under the ISO standard, almost anything can be an object class. In application, certain choices must be made. A test that was often used in trying to isolate the object class for our purposes was to answer the question "What is being counted here?" For example, we were often tempted to identify "industry" as an object class. However, statistical agencies do not report meaningful statistics on the number of industries (i.e. in 2004, there were 212 active industries in Canada). Rather, we generally report on the number and size of businesses, classified by industry. We therefore only ever considered "industry" as a property, usually assigned to an object class such as establishment or business location. Another choice was made to consider the "Economy" as an object class, which was used extensively for variables encountered in national accounts. Here, the test of "what is being counted here?" is not followed. National statistical agencies do not report statistics on the number of economies. However, across countries, there exists what can be considered a population of economies, each with its own characteristics of size, composition and behaviour. We have therefore defined many data elements with Economy as an object class, such as Value of GDP of Economy.

ISO/IEC 11179 is a very flexible standard. As mentioned earlier, almost anything can be considered as an object class by adding more and more properties to a fundamental object class. Considering the whole definitional space available to include object class – property – value domain, it is possible under the standard to pack all the meaning of a data element into the object class and to reduce the number of properties down to one (Occurrence of), with a value domain of Yes or No. This is in fact how categorical variables are represented as "dummy variables" in statistical analysis processes and software. For example, the object class "French language speaking employed person" can be defined, with the property of "Occurrence" and a value domain of Yes or No. The result, however, is a very large number of data elements (in this case, one data element for every language of interest), making their administration more difficult and the task of finding data more onerous for users. Instead, we have chosen to consider Employed person as the object class, Language spoken as the property and Language categories as the value domain. We therefore get the data element Category of Language spoken of Employed person. Our choice of limiting the number of object classes pushes more meaning into the property and value domain definitional space. We believe this balanced approach improves the administration of the data elements and ease of search and promotes the reuse of elementary definitional components, thus supporting data harmonization.

5. PROPERTIES

Properties are simply the characteristics of interest of the unit of observation (object class). Having defined the object class, it is relatively straightforward to identify which characteristics are being measured. There are a few

special conditions, however, for which consistent rules were developed in application. These cover the special properties of time, place and occurrence, and the use of combined or compound properties. Each will be discussed in turn.

Every observation of a statistical unit refers to the condition of that unit as of a particular point in time or period of time. This is often called the reference period or reference date of an observation. One might consider this time reference as a property or characteristic of the statistical unit and therefore use it in the construction of the variable. While it is without doubt essential for the correct interpretation of the data, in our application of the standard, time was not considered as a property of the object class. Rather, we consider it to be a property of the observation itself, attached to the process of observation rather than to the intrinsic meaning of the data element. The information on reference period is therefore stored in the Administration and identification region of the meta-model, rather than in the data element concept region.[7] In addition to creation date, effective date and last change date, which are proposed in the standard for time references, we created a timeframe type called "Reference period", which contains the relevant time information for a given observation or a set of observations. Usually, observations are made in the context of a given instance of a survey (for example the July 2005 instance of the monthly Labour Force Survey) and all observations in that set inherit the same reference period.

Another property that applies to most, if not all, observations is some kind of geographic reference. However, unlike time, this property is not inherited from the observation process itself. It truly is a property of the object class. We therefore have identified and defined a geographic location property, which in most cases is represented by the name of a geographic location. We have therefore consistently structured a data element if the form of Name of Geographic location of Object class to deal with the "geography" variable. This Name representation can then use a variety of value domains of geographic names, such as the Standard Geographical Classification 2001 Canadian Provinces and Territories .

We have also adopted a standard convention regarding a third special property, which we call the "occurrence" property. This property is used when the data element being documented is the result of a statistical operation, such as compilation. A compilation process produces a new data element, an incidence count of the unit observation for a given value in a value domain. This data element takes the form of Count of Occurrence of Object, where Count is the representation class, Occurrence is the property and Business Location is the object class. This Occurrence property has been consistently defined therefore wherever incidence counts are reported.

This example also illustrates another convention adopted when identifying data elements. In our application of the standard, when statistical data are presented in a table, data elements are not identified and defined for each cell in the table. The design and labeling of the table convey information that does not need to be stored in the data element definition. Only the dimensions of the table need to be defined as data elements and documented

At present, 220 properties have been loaded in the production database, with another 280 awaiting review and validation in our staging area. It appears as if the specification, naming and definition of about 500 properties will be sufficient to cover all statistical data published in CANSIM.

6. REPRESENTATION TYPE

The metadata for a data element is not complete without describing its representation. Indeed, this is the form that the data element will take in a data file or statistical table. In our application, a representation type is first identified. This describes the specific form of the representation and is used in the data element name. Terms such as count, value and number are used as representation types in the case of data elements with non-enumerated value domains, known in statistical terminology as continuous variables. Value of Expenses of Business location is an example of such a data element, using the representation type "Value". In the case of data elements with enumerated value domains, representation types such as name, type and category have been used, as in the data element Category of Age of Person. In the end, we found that a relatively small set of about 25 representation types was sufficient to cover all of the statistical output of the Agency.

[7] ISO/IEC 11179-3 (second edition), p. 32

7. DATA ELEMENTS

With these elementary building blocks having been defined, it is possible through the combination and permutation of object class by property by representation type to specify, name and define all data elements produced by a national statistical agency. This is where the consistent application of ISO/IEC 11179 could yield major harmonization gains across national statistical agencies. The list of object classes and representation types presented in this paper is almost surely applicable to all national statistical agencies. Harmonization efforts would therefore concentrate on common names and definitions of properties, which is an achievable goal. This would allow users at least to locate common data elements across national statistical agency data holdings and be secure in knowing that underlying definitions are the same. There are, however, considerable, and in many cases warranted, disharmonies in the value domains, or classifications, used to represent these data elements.

8. VALUE DOMAINS

Depending on the representation type, data elements have value domains that are enumerated or non-enumerated. Non-enumerated value domains are used to represent continuous variables, variables that can assume any numerical value within a range. For example, the non-enumerated value domain for the data element Value of income of person might be the set of integers between 0 and infinity. All that is needed to understand and interpret such a value is to know the unit of measure (i.e. Canadian dollars), and the precision (for example, two decimal places). Rather more information is required, however, to interpret values taken from an enumerated value domain. An enumerated value domain is a set of categories, represented by codes or labels, or both, each having a meaning unrelated to its actual value. The number 325410 for example is a NAICS code meaning Pharmaceutical and Medicine Manufacturing. It is impossible to know this without reference to metadata. In the standard therefore, enumerated value domains are made up of pairs of permissible values (codes) and value meanings (labels), which can also have a definition.

Value domains can also be considered standalone building blocks, which can be associated with appropriate data elements as required. Managing, registering and maintaining these value domains is in fact a common task of national statistical agencies, where they usually take the form of statistical classifications. In our application of the standard, we have re-created statistical classifications from value domains specified according to the standard. As a "classification" entity does not exist in the standard, a number of conventions were therefore developed for this purpose.[8]

First, every value domain was given a top value domain, containing only one permissible value and value meaning, which is the parent of all subordinate permissible values. This value domain is a place holder or organizational device designed to be the container for the classification of interest. This value domain is given the name of the classification that it is intended to represent (for example, NAICS Canada 2002). Under this value domain, one or more value domains are hierarchically identified. Each value domain is a set of permissible values, with associated value meanings, which are mutually exclusive and exhaustive of the universe of observations to be classified. Each value domain is assigned a level within the hierarchy. Every permissible value is assigned to a parent permissible value from a higher level value domain and its order among siblings is recorded. Each permissible value can be the child of one and only one parent permissible value and is thus exclusive in aggregation. With these conventions, the full structure of any classification can be reconstructed.

Original classifications, which might be standard classifications (and recorded as such in the registration status – see Administration and identification region of the standard), and their variants are treated this way consistently in the IMDB. The original classification is considered an "umbrella" value domain and is flagged as such in the metadata registry. In this way, potential targets for future harmonization are easily identified.

[8] Part 2 of ISO/IEC 11179 deals with classification, but this relates to the classification of data elements and their constituent building blocks in a metadata registry for ease of organization and search, which will be covered in section 9 of this paper, not the classification of observations in an enumerated value domain, which is the issue here.

Some value domains have a large number of variants. For example, Statistics Canada currently publishes data according to 13 different variants of the North American Industry Classification. Some value domains also change over time, but these are considered as versions rather than as variants. For example, NAICS Canada 1997 was the standard classification for type of industry. It had many variants under the same umbrella. When the original was replaced by NAICS Canada 2002, this was considered a new version of the same value domain. Similarly, any variants of NAICS Canada 1997 that were updated under to the 2002 version were considered new versions of these variants. Continuing with this example, there are also other classifications used for type of industry, for example the International Standard Industry Classification (ISIC) and the European industry classification (NACE). These are all related and this is represented in the model by having the value domains making up all of these related classifications use value meanings drawn from a common pool of value meanings, which is the industry conceptual domain. Conceptual domains then, in our application of the standard, are containers of value meanings, which are re-used in value domains that represent data elements built from the same representation class and property (for example, type of industry).

At present, there are 1145 value domains loaded into the IMDB, which are used on their own or in conjunction with others to form 550 classifications.

9. ORGANIZING DATA ELEMENTS

The application of the ISO/IEC 11179 standard for documenting variables published by Statistics Canada has yielded approximately 900 named and defined data elements. To help users navigate a list with this many elements, some kind of organizing framework is desirable. Data elements can be organized according to their constituent parts, for example, all data elements that use the same object class, or the same property, or that are represented using the same value domain. Such a presentation might be useful to users. However, the standard also makes provision to associate classification terms with each administered item in a registry, such as a data element. In Statistics Canada, each data element has been assigned one of 27 high level topics.

10. CONCLUSION

Using ISO/IEC 11179 to guide the task of documenting data elements at Statistics Canada has provided a rigorous framework for identifying, naming and defining a vast amount of statistical information with a relative economy of metadata terms and constructs. The power of the standard lies in the way complex data constructs can be decomposed into components that can be reassembled and reused to provide complete definitional and contextual metadata. This granularity also provides a powerful avenue for standardization and harmonization as agreement can be more easily achieved on the identification and definition of specifically delimited components such as object classes and representation classes than on complex ill-formed data constructs. Overall, this standard has been instrumental in helping the Agency achieve a long sought after goal of providing data users with an inventory of the data elements it produces and associated metadata.