Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2003: Challenges in Survey Taking for the Next Decade

2003

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2003
Challenges for Survey Taking for the Next Decade

CONCEPTUAL MODEL FOR THE DEFINITIONAL METADATA OF A STATISTICAL AGENCY

Kaveri Mechanda, Paul Johanis and Michael Webber[1]

ABSTRACT

This paper presents the Framework for Subject Matter Standards (FSMS), a conceptual model for the definitional metadata of a Statistical Agency. This framework shows a global picture of the *statistical units* and *variables* used in Statistics Canada. It was created to assess where new standards development is needed. It also provides a framework for organising the metadata and structuring definitions in the Statistics Canada's Integrated Meta Data Base (IMDB). This model borrows concepts from object-oriented modelling and translates the conceptual metadata of Statistics Canada to the ISO/IEC 11179 framework. This paper explains the work done on the FSMS and includes charts showing the basic model.

KEYWORDS: ISO/IEC 11179; Metadata; Statistical Units; Variables.

1. CONTEXT

The need for metadata standards, particularly definitions that are consistent among the myriad subject matter of a statistical agency is more urgent today that it has ever been. The increasing reuse and reprocessing of data since machine-accessible data became widely available in the 1960s has been driving the need for more and better quality metadata (Norbotten, 1993). More recently, the availability of diverse types of data and information on the Internet means that it is possible for users to view and compare data from one subject matter area with that of another. This has focused the need for consistent definitions across subject matter areas, over time and between agencies. Also, with retirements and changes in the staff within a statistical agency, well-organised metadata becomes an important means for knowledge transfer. Rapid changes in technology necessitate a conceptual model that is independent of technology. The wide use of statistical data, turnover of staff and changes in technology drive the current need for a conceptual metadata model within the statistical agency.

Statistical agencies create statistical metadata – metadata that is needed for the proper production and usage of statistical data. The users of statistical metadata have various needs for the interpretation and use of the metadata. Sundgren identifies three principle types of information made available with the data (Struijs 1999, Sundgren 1993).

These are:

- Pragmatic information which informs users about the purpose and use of the data. Information on how to use a product could be an example of this.
- Semantic information which informs users about the meaning and content of data. Definitions of variables are examples of this.
- Syntactic information which informs users about the physical and technical aspects of the data. This includes data representations and data processes.

Definitional metadata is situated for the purpose of the Framework for Subject Matter Standards as being a type of semantic information for statistical metadata terminology.

Statistical agencies are mandated to produce data for the public good. Concepts then, must be tuned to the public interest. The approach presented here models the metadata from the data user's point of view or more specifically in

[1] Kaveri Mechanda, Senior Analyst, Geography Division, Statistics Canada, Ottawa, Canada, K2E 7P8, Paul Johanis, Director, Standards Division, Statistics Canada, Ottawa, Canada, K2E 7P8.

terms of the data that are disseminated rather than in terms of the data that are part of the data collection process in a statistical agency. The reasoning is that definitional metadata should be user driven but also, co-ordinated with the data collection process of a statistical agency.

The focus on the creation of statistical metadata by statistical agencies as well as national and international organisations has resulted in concepts and framework for the same. Although not specifically created for statistical metadata, *the ISO/IEC 11179 Part 1, Framework for the specification and standardization of data elements*, has been an influence and has been used by statistical agencies. Object-oriented modelling has increasingly gained popularity in modelling systems. This is now a mature modelling tool and is insightful in modelling statistical metadata. Much work has been done in the conception of the storage and processing of metadata, particularly the metadata pertinent to data collected by means of surveys. Less work has been done in the conception (globally) of definitional metadata for data dissemination.

Statistics Canada is in the process of populating the Integrated Meta Data Base (IMDB) with information about surveys as well as with the associated definitional metadata. This metadata registry is updated on a daily basis with metadata for data that are published by Statistics Canada on that day. It provides users with access to the metadata directly from the Internet. Since this database is expected to store all of the metadata for data published on the Internet, a consistent conceptual framework is needed to support and standardise its definitional metadata.

The objective of Statistics Canada's Policy on Standards is to standardise definitional metadata. This policy deals with reviewing, documenting, authorising and monitoring the use of standard definitions for the subject-matter concepts, *variables* and *classifications* used as well as the *populations* and *statistical units* to which they apply (Statistics Canada, 1988). Standards are created under the direction of the Policy on Standards. To proceed with a systematic and global approach to creating new standards, a conceptual model is needed for all of the definitional metadata of Statistics Canada.

The Framework for Subject Matter Standards (FSMS) provides a conceptual framework for the definitional metadata of a statistical agency. It supports the IMDB as well as the creation of new standards. The FSMS will be populated with the metadata from the IMDB. Then, the metadata in the FSMS will be evaluated in terms of where harmonisation and structuring of definitions is needed, and where standard classifications need to be developed. Newly created standards will be registered and made available through the IMDB. In this way, through the IMDB, the FSMS will provide a systematic way of realising the objective of standardising definitional metadata.

2. WHAT IS THE FRAMEWORK FOR SUBJECT MATTER STANDARDS?

The Framework for Subject Matter Standards (FSMS) was created to provide a consistent and comprehensive conceptual model for the definitional metadata used in Statistics Canada. The framework is expected to show where definitional metadata standards are missing in Statistics Canada. It models definitional information – definitions that people need to know rather than information that machines need (syntactic information). Providing a comprehensive view of the definitional metadata is considered a first step in the harmonisation of the metadata.

The Framework was created with the following principles in mind:

- Effectiveness – the model should be usable and revealing of the metadata used in Statistics Canada. For example, it should be practically applicable in the context of the IMDB and it should reveal areas where further standardisation work is needed.
- Comprehensiveness – the model should be able to include all the *statistical unit*s and concepts that are within the scope of information published in Statistics Canada.
- Efficiency – the model should be as simple and as clear as possible.

2.1 Influences on the Framework

The following influence the structure of this framework:

- Statistics Canada's Policy on Standards
- Object-Oriented Modelling
- ISO/IEC 11179

Ways of systematically ordering information and metadata as found in the object-oriented paradigm and ISO/IEC 11179 have been applied to the organising of the statistical metadata and information. Object-oriented modelling was designed for use in modelling system applications and ISO/IEC 11179 was originally designed for data exchange. Ideas were borrowed from both these ways of modelling in creating the framework. Notes on the equivalencies between the terms used in ISO/IEC 11179 and object-oriented modelling are presented in Appendix A. Since Statistics Canada's IMDB follows the ISO/IEC 11179 model, co-ordinating the equivalencies of terms used within Statistics Canada with the standards and the object-oriented modelling concepts is important.

The FSMS models the definitional metadata identified in Statistics Canada's Policy on Standards. It currently models *statistical units* and *variables*. The framework will be extended to cover *classifications* and *populations*.

2.2 Content of the Framework

The organisation of the metadata in the model is deliberately influenced by the way statistical subject matter is viewed and organised within Statistics Canada. This was done with the intent that the framework should be based on accepted ways of viewing the subject matter and it should be technology independent.

The Framework currently consists of a chart for *statistical units* and a chart for *variables*. The chart for statistical units consists of four views that divide the model into four sections. These four sections are based on different focuses of study of the subject matter. It also defines different types of statistical units providing structure to the definition of the relationship between statistical units. The chart for variables shows how variables are (or could be) organised.

The scope of this modelling exercise is currently limited to the information that is authored and published on a regular basis by Statistics Canada. For example, data authored by other agencies that are hosted on the Statistics Canada web site and information presented in analytical articles has not been examined for inclusion in this modelling exercise. However, data collected by other agencies that are processed and published by Statistics Canada (i.e. some administrative and environmental data) are included in this exercise.

3. STATISTICAL UNITS AND POINTS OF VIEW WITHIN THE FRAMEWORK

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" (Statistics Canada, 1988). The organisation proposed by the framework for these *statistical units* is presented below.

3.1 Points of View within the Framework

To comprehensively cover the data and information published by Statistics Canada, different views are accommodated within the framework. Four *macro statistical units* were chosen to provide four different views within the model:

- People
- Economy
- Environment
- The State

These *macro statistical unit*s closely resemble the commonly used pillars by which data have been disseminated by Statistics Canada. They also provide four different perspectives in that, the "People" view provides the perspective of persons – a more sociological view; The "Economy" view provides the perspective of business units and economic units – a more economic view; The "Environment" view is intended to provide a natural capital perspective; "The State" is intended to provide a window on governance in Canada. In order to picture the activities of governance in a holistic way, "The State" section includes private provision of state mandated services.

These *macro statistical units* may be viewed as the basis for which *statistical unit* data are published or they may be viewed as aggregations of *fundamental statistical units*.

3.2 Fundamental Statistical Units and Stereotypes

There was a conscious effort to make the model simple and robust by limiting the number of *statistical unit*s. An object-oriented modelling approach was taken to typify *statistical unit*s. An object is defined as a "concept, abstraction or thing with crisp boundaries and meaning for the problem at hand" (Rumbaugh, Blaha, Premerlani, Frederick and Lorenson, 1991). In this exercise, the number of objects used to characterise the data published by Statistics Canada is kept to a minimum. This was done by first identifying a small number of *fundamental statistical unit*s. Commonly used derivations (e.g. subsets, groupings) of *fundamental statistical unit*s were also modelled.

*Fundamental statistical unit*s are defined as those that are not types of any other *statistical unit* and can not be derived as grouping of any other *statistical unit.*

The *fundamental statistical unit*s were stereotyped to reveal the differences in how these units behave. The *fundamental statistical units* were divided in the following way:

- *Statistical units* that operate and whose operations are reported on by Statistics Canada are stereotyped as "*agents*". In the "People" view "Person" can be considered as an *agent*.
- *Statistical units* that represent the actions of (or by) *agents* as reported by Statistics Canada are stereotyped as "*events*". *Events* are defined as occurrences that are discrete in time (occur in time period) and finite (can be counted). In the "People" view "Birth" can be considered an *event.*
- Other *statistical units*, reported on by Statistics Canada, which are generally either produced or managed by *agents*, are stereotyped as "*items*". In the "Economy" view "Product" can be considered as an *item.*

Examples of *statistical units* and their stereotypes are shown in Table 1, Chart for Statistical Units. Identification of the *agents* within each of the four *macro statistical unit* views reveals the scope of each view. Each view accommodates metadata from the point-of-view of its Agents. Agents of any *macro statistical unit* may interact with agents of any other *macro statistical unit.*

3.3 Commonly Used Derivations of Fundamental Statistical Units

Fundamental statistical units were identified as a means of keeping the number of units to a minimum. However, many *statistical units* are so commonly used that they need to be incorporated in the framework. These are all *statistical units* that can be derived in some way from the *fundamental statistical units*. Three common derivations of the *fundamental statistical unit* are:

- *Subclass statistical units* based on inherent property. Examples of this are "Men" or "Women" used as *statistical units* for which "Person" data are reported.
- *Roles* that the *statistical units* may assume. Examples of this for the Person *Fundamental statistical units* are "Student", "Mother" and "Employee". Examples of this for the "Business Unit" are "Producer", "Employer", and "Consumer". *Roles* differ from *subclasses* 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.
- *Aggregations* or *aggregation hierarchies* of *fundamental statistical unit*s. For example, "Census Family" is an *aggregation* of "Persons". "Enterprise", "Company", "Establishment" and "Business Location" form an *aggregation hierarchy* for "Business Units".

Table 1
Chart for Statistical Units

Macro **Statistical Unit**	**Fundamental Statistical Units**	**Some Commonly used derived Statistical Units**
People (of Canada)	Agents • Person Events • Birth • Death • Marriage • Divorce • Immigration • Emigration	**Persons** Common Subsets: • Age based: **Children, Youth, Adults, Seniors** • Age and sex based: Men, Women Common Roles: • Student • Employee • Traveller • Consumer • Mother • Father Common Aggregations: • **Household** • **Census Family** • **Economic Family**
Economy (of Canada)	Agents • **Institutional unit** • Business unit Events • Economic transaction • Bankruptcy Items • Product • Building-dwelling • Job • Economy	**Institutional Units** Common Subsets: • **Non-financial corporations sector** • **Financial corporations sector** • **Government sector** • **Non-profit institutions serving households sector** • **Households sector** Common Roles • Producer • Employer • Consumer Business Units Common Roles • Producer • Employer • Consumer Hierarchies of Aggregations: • **Enterprise** • **Company** • **Establishment** • **Business Location**
Environment (of Canada)	Agent • Nature Events • Environmental accident • Ecosystem quality Items: • Agricultural resources • Forest resources • Marine resources • Water resources • Energy resources • Mineral resources • Wildlife resources • Air and atmosphere resources	Environmental Accidents, Common Subsets: • Dangerous goods accident • Forest fire Ecosystem Quality, Common Subsets: • Ground level ozone exceeds objective Mineral resources, Common subsets: • Nickel • Potash Energy Resources, Common subsets: • Natural gas • Crude oil Forest resources, Common subsets: • Timber
The State	Agents • The State Events • Permit • Claim • Admission • Offence • Charge	The State, Common subsets: • Government body • Educational institution • Justice institution • Health institution Permit, Common subsets: • Building permit Claim, Common subsets: • Unemployment insurance claim

The Chart for Statistical Units shows how *statistical units* are organised in the Framework for Subject Matter Standards (See Table 1). The first column (on the left) lists the four views provided by the *macro statistical units*: The People, The Economy, The Environment and The State. The next column shows the *fundamental statistical units* associated with each view, with examples for the stereotypes *agents, events* and *items*. While the chart shows how the framework organises statistical units, it does not yet show a comprehensive view of all the *statistical units*. Working with the data in the IMDB is expected to expand and improve the content of the model. *Statistical units* for which standard definitions exist are shown in bold; standardisation work needs to be done where the *statistical units* are not in bold.

4. VARIABLES WITHIN THE FRAMEWORK

*Variable*s are organised in this framework in a hierarchy with *variable* at the bottom of the hierarchy, *variable concept*s in the middle and *variable group* at the top of the hierarchy. These terms are explained below. An example of a *variable group* would be a general subject of study such as "Ethno-Cultural Characteristics". A *variable concept* associated with that *variable group* would be "Language". *Variables* associated with that concept for a person would be "Mother Tongue", "Language Spoken at Home" and "Official Language(s) Spoken".

4.1 Variables

Variables are defined by Statistics Canada's Policy on Standards as a *specific characteristic or attribute of a statistical unit* (Statistics Canada, 1988). The United Nations Glossary of Classification Terms further elaborates this as follows:

A characteristic of a unit being observed that may assume more than one of a set of values to which a numerical measure or a category from a classification can be assigned (Expert Group on International and Social Classifications).

In the context of describing a population, a *variable* consists of the representation of the *variable concept* with a classification or unit of measure. The *statistical unit* for which the *variable* is observed is not present in the name of the *variable* in this framework. However, the *statistical unit* is implicit.

4.2 Variable Concepts

Properties of *statistical units* are observed and reported on by Statistics Canada. These properties are considered as *variable concepts*. *Variable concepts* are abstract in that they are **not** associated with any classification or unit of measure. From the object-oriented design perspective they may be considered properties of an object class for the *statistical unit*.

Some *variable concepts* (properties) may be considered to be common to all *statistical unit*s. Occurrence is a property of every *statistical unit*. (In the sense that it is observed, therefore, it occurs.) This is the *variable concept* when simply the number (rate, percentage) of a unit is reported. Location and time are the other two *variable concepts* that are common to *statistical units*. Other properties of *statistical units* are often cross-classified by time or location in tabular presentations of aggregated statistical data.

4.3 Variable Groups

*Variable group*s are a collection of *variable concept*s. *Variable concepts* may be grouped in different sets in order to provide different perspectives on the phenomena that are represented by *variable concept*s. These groups do not have to be mutually exclusive. For example, Language may be considered as a Human Capital Characteristic and an Ethno-Cultural Characteristic. These groupings may be re-created to fit a particular use.
This way of grouping variables can also be co-ordinated with the way metadata are searched on the Statistics Canada web site. Statistics Canada's web site has a "Browse by subject" section that enables the user to search the web site by terms in the Theme List. These are generally terms such as "Trade", "Agriculture", "Health", "Justice,"

etc. For searching on the web, *variable concepts* could be put in sets of *variable groups* represented by the titles in the Theme List. This theme way of classifying subject matter may also be applied across the *statistical units* within each *macro statistical unit* view. For example, for the theme "Agriculture", People *statistical units* (persons) are "farm operators", Economic *statistical units* (businesses) are "farms" and the Environment *statistical units* (Land and soils) may be effected by the farm operation, and The State *statistical units* (Educational institutions) may offer courses in "Agriculture".

5. APPLICATION OF THE FRAMEWORK FOR CREATING DEFINITIONS

The FSMS is designed to organise definitional metadata and assess where new standards development is needed. The framework that has been developed also suggests a structured approach in the development of definitions. Well-structured definitions for the metadata make it possible to better understand and use the data. Such definitions are also easier to transcribe into programming code.

Examples of some structure rules that may be required in definitions of *statistical unit*s are:

- For a *statistical unit*, define the relationship of this *statistical unit* to other *statistical unit*s. For example the following questions could be asked:
  - Is this a *fundamental statistical unit*?
  - If it is not a *fundamental statistical unit*, what *fundamental statistical unit(s)* is it derived from? What is its stereotype?
- For *a statistical unit*, what *variable* values need to be part of the definition?
  - For example, what is the Age range that classifies a Person as belonging to the *subclass* "Senior"?
- For *statistical units* that are *roles*, the conditions for assuming the *role* and ending the *role* should be specified.
  - For example, is a person considered a student only for the time period when that person is enrolled in a certain minimum number of courses?

Given these rules, a question that could be asked when creating the definition for "Child" would be: "Is this based on Age (*variable* value range) or based on relationship to another *statistical unit* (Parent)? These rules for definitions may make them easier to think through and construct.

Further development of the framework will lead to the delineation of rules for *variables*, *classifications* and *populations*.

Providing a structured way to think about definitions for the subject matter of a statistical agency could make it easier to create these definitions and to harmonise the definitions with each other within the Framework for Subject Matter Standards.

6. FUTURE WORK

The Framework for Subject Matter Standards provides a global view of the definitional metadata of Statistics Canada. As the charts in the framework are filled in with more metadata, they are expected to show where work needs to be done to create new standards. The FSMS provides a consistent and comprehensive framework for the definitional metadata of the IMDB.

Further work along the lines that are suggested by the FSMS would involve the following:

- Expanding the framework to include and define all the *statistical units* and *variables* for which data are published on a regular basis by Statistics Canada.
- Creating new *classifications* where they are needed.
- Reviewing the definitions of the *statistical units* and *variables* and harmonising these across subject-matter areas, over time and between agencies.
- Using the Framework as a "Road Map" to search for definitional metadata on the Internet.

The FSMS is also being considered for use in other applications such as the work related to assigning metatags for the Statistics Canada web site.

The FSMS itself needs to be further developed to include *classifications and populations* as well as possibly delineate *variables* in greater detail. Work being done for the update of the IMDB, particularly in elaborating *classifications*, will provide further input for the development of the FSMS (Johanis, Brooks, Dunstan and Lévesque, 2003)

The need for definitional metadata standards that are consistent among the subject matter of a statistical agency is more urgent today that it has ever been. The work on the FSMS is critical to organise definitional metadata as well as to make the metadata coherent, interpretable and easy to find.

APPENDIX A
Equivalencies of Terms

Terms used in the ISO/IEC 11179 framework and object-oriented modelling, are similar and yet different. The chart below shows the equivalencies of various terms to that of the Framework for Subject Matter Standards.

Object-Oriented Modelling **(Definitions are from books about object-oriented modelling)**	**Framework for Subject Matter Standards**	**ISO/IEC 11179** **(Definitions are from the ISO/IEC 11179 publications.)**
Object …a concept abstraction or thing with crisp boundaries and meaning for the problem at hand…objects have identity and are distinguishable (Rumbaugh, Blaha, Premerlani, Frederick and Lorensen, 1991)	Statistical Unit The unit of observation or measurement for which data are collected or derived (Statistics Canada, 1988). A statistical unit is specified in terms of its relationship to other statistical units and/or in terms of variable value ranges.	Object Any part of the conceivable or perceivable world (ISO/IEC, 1999).
Object Class …describes a group of objects with similar properties (attributes), common behaviour (operations), common relationships to other objects, and common semantics,…(Rumbaugh, Blaha, Premerlani, Frederick and Lorensen, 1991). Class A model element used to model things, both intellectual and physical. A class is a type… (Ericksson and Penker, 1998).	Population The set of statistical units to which a data set refers (Statistics Canada, 1988). Population is specified in terms of space (geographic area) and time period.	Object Class A set of objects. A set of ideas, abstractions, or things in the real world that can be identified with explicit boundaries and meaning and whose properties and behaviour follow the same rules…Object classes can be formed by combining two or more other object classes (ISO/IEC, 1999).
Property An abstract state of the object (Norbotten, 1993).	Variable Group A general or abstract idea that expresses the social and/or economic phenomena to be studied or observed. An abstraction that can be used for the modelling of a set of variable concepts. This concept is independent of the statistical unit.	Property: A particularity common to all members of an object class... (ISO/IEC, 1999).
	Variable Concept A general or abstract idea that expresses the social and/or economic phenomena to be measured. Abstract property of the object class. It is independent of the representation. Variable concepts are terms from the subject-matter taxonomy for a variable group. They may be greater in scope than a data element concept.	Data Element Concept A concept that can be represented in the form of a data element described independently of any particular representation …A combination of an object class and a property is a data element concept (ISO/IEC, 1999).
Attribute …a data value held by the objects in a class…Each attribute has a value for each object instance…each attribute name is unique within a class…different object instances may have the same or different values for a given object class (Rumbaugh, Blaha, Premerlani, Frederick and Lorensen, 1991). This implies mapping to a representation	Variable A specific characteristic or attribute of a statistical unit (Statistics Canada, 1988).	Data Element A unit of data for which the definition, identification, representation, and permissible values are specified by means of a set of attributes…a data element can also be seen to be composed of two parts: a data element concept and a representation (ISO/IEC, 1999). *A variable applies to statistical units and the name of a variable does not contain the name of a statistical unit. Because of this, the Generic Data Element, which consists of the property and the representation, may be considered equivalent to the variable.*

REFERENCES

Colledge, M.J. (1999), "Statistical Integration Through Metadata Management", *International Statistical Review*, Vol.67, No. 1 pp.79-98.

Eriksson, H-E. and M. Penker (1998), *UML Toolkit*, John Wiley and Sons, Inc.

Expert Group on International Economic and Social Classifications, *United Nations Glossary for Classification Terms*. Available at: http://unstats.un.org/unsd/class/family/glossary_short.htm

Gilbert, N. G. (1981), *Modelling Society*, George Allen & Unwin (Publishers) Ltd.

Grossmann, W. (1999), *Statistical Metadata Tutorial*, available at: http://www.univie.ac.at/dac/papers/Tutorial_gesamt.pdf

ISO/IEC (1999), ISO/IEC 11179-1 *Framework for the specification and standardisation of data elements*, First Edition.

Johanis, P., (2000), "Statistics Canada's Integrated Metadatabase, Current Status and Future Plans", paper presented at the Work Session on METIS, Statistical Commission and Economic Commission for Europe Conference of European Statisticians.

Johanis, P. (2002), "Assessing the Quality of Metadata: The Next Challenge", paper presented at thethe Work Session on METIS, Statistical Commission and Economic Commission for Europe Conference of European Statisticians.

Johanis, P., B. Brooks, T. Dustan, and J-S. Lévesque, (2003), "Statistics Canada's Implementation of the Data Element Model", paper for the Metadata Registries Open Forum.

Muller, R. J. (1999), *Database Design for Smarties, Using UML for Data Modelling*, Morgan Kaufmann Publishers.