WP.7
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD) STATISTICS DIRECTORATE

Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS)
(Geneva, 3-5 April 2006)

Session (ii) Metadata Concepts, Standards, Models and Registries

METADATA STANDARDS AND THEIR SUPPORT OF DATA MANAGEMENT NEEDS [1]

Invited Paper

Submitted by Statistics Canada[2], Canada, and the Bureau of Labor Statistics[3], United States
Session (ii) Metadata Concepts, Standards, Models and Registries

I. INTRODUCTION

1. For many years, metadata management has been an important concern in national and international 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 and automate 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, Kent, 1998; Kutin and Arnic, 2004; Dunnet and Merrington, 2006)

2. Several things need to be understood and developed before metadata management and metadata driven systems 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).

---

[1] The opinions expressed in this paper are those of the authors and do not necessarily reflect the official policies of Statistics Canada or the Bureau of Labor Statistics.

[2] Prepared by Paul Johanis at Statistics Canada. Contact paul.johanis@statcan.ca.

[3] Prepared by Daniel W. Gillman at the Bureau of Labor Statistics. Contact gillman.daniel@bls.gov.

3. 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.

4. Most situations require some amount of modeling work. 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 often the case.

5. Where does one look for appropriate models? There are 4 possible answers: other statistical offices, commercial software vendors, 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; Sundgren, 1995). 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 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.).

6. Standards and other statistical offices seem to be the best sources for finding appropriate metadata models, and we will analyze several metadata schemes, which arose from these 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 these common constructs.

7. The paper is organized into several sections. We begin with a section on the theory of terminology. This provides a framework for commonality. Next, a discussion of statistical data based on terminology theory is provided. Then, we show that the ISO/IEC 11179 standard is an implementation of the theory. Therefore, the standard is common to all descriptive frameworks for statistical data. The Corporate Metadata Repository (CMR) model, an extension of the ISO/IEC 11179 standard into the statistical survey domain, is discussed. Following, is a description of the most important constructs for each of five metadata schemes: Common Warehouse Model (CWM), Data Documentation Initiative (DDI); eXtensible Business Reporting Language (XBRL), Neuchâtel Variables and Classification models; and Statistical Data and Metadata Exchange (SDMX). Finally, a comparison between the models and a framework for using the models together in a statistical office are provided.

II. THEORY OF TERMINOLOGY

A. Basic Definitions

8. Terminology is the study of concepts and their representations in special language. It is multidisciplinary, drawing support from many areas including logic, epistemology, philosophy of science, cognitive science, information science, and linguistics. Work in the area dates all the way back to the ancient Greek philosophers.

9. 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:

- **object** - something conceivable or perceivable
- **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.)*
- **characteristic** - abstraction of a *property* of a set of *objects*
- **essential characteristic** - *characteristic* which is indispensable to understanding a *concept*
- **delimiting characteristic** - *essential characteristic* used for distinguishing a *concept* from related *concepts*
- **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)
- **generic concept** - *concept* in *generic relation* to another that has the narrower *intension*
- **specific concept** - *concept* in *generic relation* to another that has the broader *intension*
- **generic relation** - relation between two *concepts* where the *intension* of one of the *concepts* includes that of the other *concept* and at least one additional *delimiting characteristic*
- **subject field** - field of special knowledge

10. 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. Signs, through which designations are represented, are left undefined, but a sign is what a person perceives and interprets as designating some concept. Basically, however, a sign is a concept whose extension is a set of perceivable objects. Examples of signs are each of the lines and dots on this page we interpret as words, letters, and punctuation. So, what we see and interpret is not really a sign, but an object in the extension of the sign. The objects **F** and **F** are in the extension of the same sign.

11. Characteristics are used in concept formation. They are abstracted from properties of objects and are used to form the intension of concepts. The objects whose properties are abstracted into the characteristics that form the intension of some concept make up its extension. Characteristics may be concepts in their own right, too. They are used in concept analysis, concept modeling, formulation of definitions, and even term formation.

12. The term *specialization* is often used to denote the creation of a specific concept in generic relation to a given, generic, one.

13. 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).

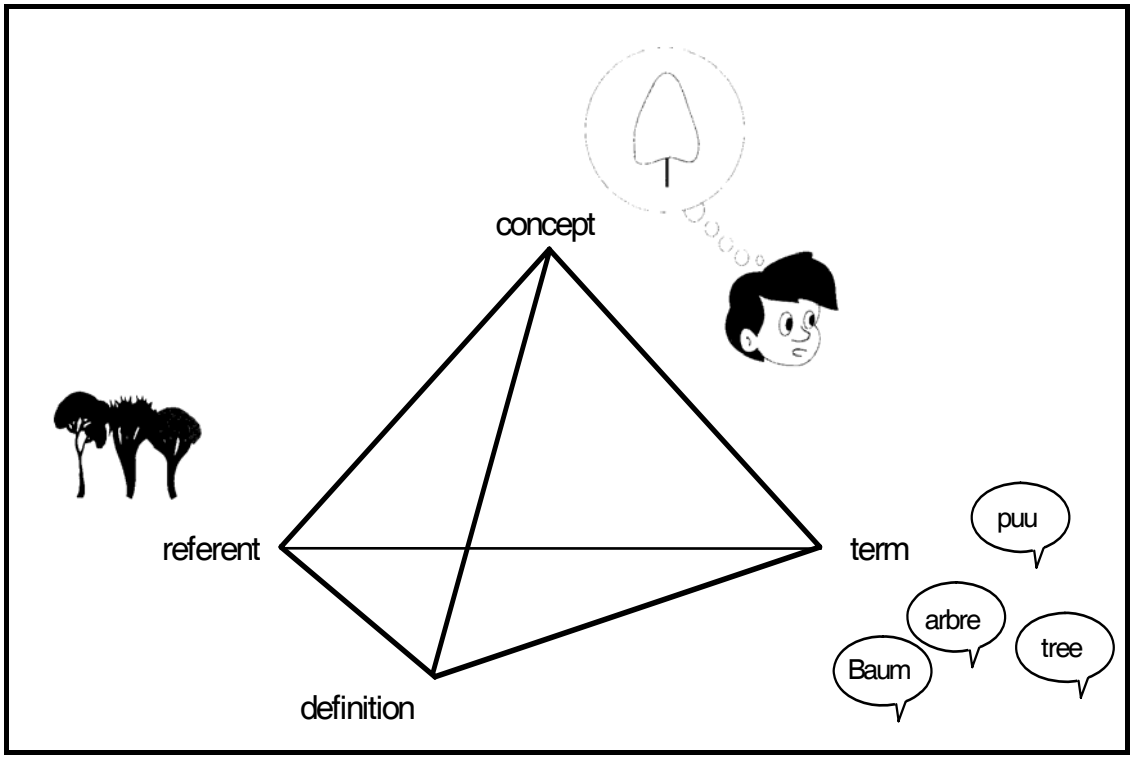


Figure 1. Relationships between referents (objects), concepts, terms (more generally designations), and definitions.

14. Figure 1 shows how terms, concepts, objects, and definitions are related. From the definitions above and Figure 1, several important observations need to be made:

- For any concept, there may be many designations (synonyms)
- For any concept, there are one or more objects in its extension
- For any concept, there may be more that one definition (especially in multiple languages)
- For each term, more than one concept may be designated (homographs)

15. 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, a robin fits more of the characteristics of a bird than a penguin does.

B. Relationship to Data

16. Statisticians view a datum as a value representing a class in a partition of a population of objects, where the partition[4] is defined for some characteristic of the population (Froeschl, Grossmann, & Del Vecchio, 2003). Here, we treat the population as a concept. Data are collected on the set of objects, the extension of the population, by measuring some characteristics of the population. For a given characteristic, the corresponding property for an object is assigned a value corresponding to one of the allowed classes in the partition.

17. In the finite case, usually for categorical data, the partition is often called a classification, e.g., sex categories. In the infinite or unbounded cases, usually quantitative data, the partition may not have a finite number of classes. Instead, the values denoting the classes come from a range of values.

18. In any case, the classes in the partition are concepts. In the sex classification example in the preceding paragraph, the classes are male and female. In the case of a range, e.g., all real numbers between 0 and 1, then we might say each value represents a probability. Because each class is a concept, then the value representing the class is a designation, in the terminological sense.

19. Therefore, a datum is a designation (Farance and Gillman, 2006). The concept associated with the designation is, at least, a combination of the population, the characteristic (as a concept) under study, and a class within the partition.

III. ISO/IEC 11179

A. Overview

20. The ISO[5]/IEC[6] 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 the semantics of data.

21. The standard is divided into six parts, each of which describe an aspect of the standard. A short description of each part follows:

- Part 1 - *Framework* -- an overview of the standard and the methodology behind data semantics
- Part - *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

---

[4] 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.

[5] International Organization for Standardization

[6] International Electrotechnical Commission

22. 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[7]. The 1st edition of the standard, published in 2000, was superseded by the 2nd. It was called *Standardization and specification of data elements*. The change in focus away from just data elements in the 1st edition necessitated the change.

23. The basic unit for describing data 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.

24. 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.

B. Implementing Terminology Theory

25. 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)

26. Without going into details about the model described in the standard and defining all the terms there, we 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).

27. Here is the list, with the terms mapped:

Terminology	ISO/IEC 11179
Concept (population)	Object Class
Characteristic	Property[8]
Partition	Conceptual Domain
Classes	Value Meanings
Designation (values)	Permissible Values

28. As stated before, the main data construct in ISO/IEC 11179 is the data element. Data elements may be abstract or implemented in some information system. Either way, a data element is a container of data (imagined or actual) where each datum has the same semantics with the possible exception that the value meaning may differ. This corresponds to the situation described in Example 1 and Example 2 below. Example 2 describes a data element with only one value.

29. Two other important ISO/IEC 11179 constructs are conceptual domain and value domain. A conceptual domain is a set of value meanings. A value domain is a set of permissible values. Since a permissible value is actually a pair consisting of a value (designation) and its value meaning (concept), the conceptual domain and value domain are closely related. If one sex codes value domain contains these permissible values

<M, male>
<F, female>
and another contains these permissible values instead

[7] Information Technology Task Force (ITTF) under ISO and IEC (http://isotc.iso.org/livelink/livelink/fetch/2000/2489/Ittf_Home/ITTF.htm).
[8] This was a most unfortunate choice. The term will be changed to characteristic in the next (3rd) edition.

<0, male>
<1, female>

then the two value domains are linked through the value meanings (male, female), which are shared.

C. Statistical Data

30. There are two senses in which the term object class is used. First of all, an object class 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. Both senses are used, and the context is intended to identify which.

31. Now, in a way that departs from traditional statistics theory a little, we say the object class may be either a general or individual concept. Two examples illustrate the ideas:

Example 1: Data element with object class as general concept (microdata)

Object class:	Adults age 16 and older in Switzerland
Property:	Sex
Value meanings:	{Male, Female}
Permissible values:	0 for Male 1 for Female

Example 2: Data element with object class as individual concept (macrodata)

Object class:	The set of adults age 16 and older in Switzerland
Property:	Proportion of females
Value meanings:	{x | 0 <= x <= 1}
Permissible values:	Real numbers between 0 and 1, with precision to 3 decimal places

32. In the case of microdata, object classes are general concepts - see Example 1. In the example, the object class is all adults age 16 or older in Switzerland. Since there are many such people (more than one), this is a general concept. However, aggregate data, or macrodata, requires an object class with one object. In Example 2, the property[9] "proportion of females" applies to the set containing all adults age 16 or older in Switzerland, not to each of the elements, i.e., each individual. The set consisting of "all adults age 16 or older in Switzerland" is a single thing. It is this aggregate that has the property "proportion of females", not the individual people. "Proportion of females" is not a property of people, "sex" is. Likewise, "sex" is a property of people, and it is not a property of the aggregate, "proportion of females" is. The point is the object classes in the two examples have different intensions, because they are different concepts, even though they are closely related. This does not mean they cannot have some properties in common, but they must have some different ones.

33. There is an exception to the rule that the object class for macrodata is an individual concept. This is the situation where the object class describes multiple instances of the same aggregate, rather than describing a single one. This arises in time series and tables, where an aggregate is reused over time or over multiple specializations. See the discussion and example in the next section.

[9]The term property is used in this paragraph and the rest of section 3 in the ISO/IEC 11179 sense, i.e., a characteristic of a concept.

D. Tables and Time Series

34. Tables are used to present cross-tabulated and time series data in an easy to read format. As the number of dimensions in a cross-tabulation goes up, so does the number of cells. It is tempting to describe each cell as a different data element; however this adds a huge burden to those responsible for recording the metadata. An easier method and equally as effective for describing tables is illustrated next.

35. Consider the simple Table 1 given below. It is a cross-tabulation of sex by age by population for all grizzly bears[10] in Jellystone Park[11].

	Population
Male	105
Age 0 - 16	60
Age 17 - 32	45
Female	95
Age 0 - 16	50
Age 17 - 32	45
Total	200

Table 1: Sex by Age by Population for Grizzly Bears in Jellystone Park

36. Table 1 is described using 3 data elements, specified as follows:

Data Element Name	ISO/IEC 11179 Constructs	Values
Sex of Bears, codes	Object class	Grizzly bears in Jellystone Park
	Property	Sex
	Value domain	<M, male> <F, female>
Age of Bears, categories of years	Object class	Grizzly bears in Jellystone Park
	Property	Age (years)
	Conceptual domain	<1, age 0 - 16> <2, age 17 - 32>
Jellystone Park bear population, counts	Object class	Grizzly bears in Jellystone Park (aggregate)
	Property	Cross-tabulated population
	Value domain	<non-negative integers; counts>

37. These 3 data elements completely describe the semantics of Table 1. All the cells are understandable from the semantics of the three data elements. For instance, to understand the cell labeled by "population of female grizzly bears age 0 to 16 in Jellystone Park", one must look at the semantics of the 'age' and 'sex' data elements to understand the conceptual domains

[10] We assume the average life span of a grizzly bear is 32 years. The data are made up.

[11] Jellystone Park is an invention of the Hanna-Barbera cartoon syndicate. The cartoon character Yogi Bear lived in Jellystone Park.

used to classify the cell. Finally, one looks at the 'counts' data element to understand what the counts mean.

38. Notice, that the object class for the third data element called "Jellystone Park bear populations, counts" is an aggregate for all the bears, but it is a general concept. Each of the cells in the table corresponds to a different Jellystone Park grizzly bear population. They are specialized populations (e.g., only females age 0 - 16), however, as stated above, one finds the semantics of the cell to understand the modification.

39. There isn't an automatic rule for deciding when an object class is a general concept versus an individual concept. It depends on use, which should be reflected in subtle differences in the properties of the concepts. When data are aggregated, the object class is clearly an individual concept. This is because a single object is under consideration. However, the aggregates in Table 1, considered as a collection, are described by an object class without recourse to some specializations, i.e., "Grizzly bears in Jellystone Park". Each object, the cells, are not described individually, they are described collectively. One could define an object class for each cell, e.g., "Female grizzly bears aged 0 - 16 in Jellystone Park" corresponds to the cell " females age 0 - 16". Then, this object class is an individual concept.

40. The same holds true for time series. There, the specialization is usually based on time. So, if we estimate the population of grizzly bears in Jellytone Park each year, then an object class that is an individual concept must include a time property, e.g., " Grizzly bears in Jellystone Park in 2006".

E. Metadata

41. When one conjures a particular object in the mind, the conception of that object is an individual concept. This is because there can be only one object in its extension, the conjured object. 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 those data are metadata. Data are only metadata when they are used to describe some object.

42. This implies all data are metadata at the point of collection!

F. Registration

43. Registration is the set of rules, operations, and procedures that apply to a metadata registry. The three most important outcomes of registration are the ability to monitor the provenance (the source of the metadata), quality of metadata, and assigning an identifier to each object described.

44. Registration also requires a set of procedures for managing a registry. The rules cover submitting metadata for registration of objects and maintaining subject matter responsibility for metadata already submitted. For actual implementations of a metadata registry, additional requirements may be necessary.

45. Provenance refers to the source of the metadata. Registration handles this in several ways:

- Naming the subject matter specialist responsible for the content of a registered item
- Naming the organization (or person) who submitted the metadata for registration
- Maintaining identifiers and version numbers

46. There are several purposes to monitoring metadata quality. The main purposes are as follows:

- Monitoring adherence to rules for providing metadata
- Monitoring adherence to rules for forming definitions and following naming conventions
- Determining whether a description still has relevance
- Determining the similarity of related data constructs and harmonizing their differences
- Determining whether it is possible to ever get higher quality metadata for some data constructs

47. Every data construct registered in a metadata registry is assigned a unique identifier. Identifiers are a means to keep track of descriptions for administration purposes, to refer to descriptions by remote users of the registry, and aid in metadata transfer between registries.

48. The registration authority is the organization responsible for setting the procedures, administering, and maintaining a registry. The submitting organization is responsible for requesting that a new description be registered in the registry. The steward is responsible for the subject matter content of each registered item. Each of these roles is described in ISO (2005).

G. Application for documenting data elements in statistical agencies

i. Introduction

49. Both the Bureau of Labor Statistics and Statistics Canada are using ISO/IEC 11179 to document data elements, or variables. The work is in different stages of development, but many of the experiences are similar. In this section, we will discuss some of the practical considerations in using the standard in statistical offices.

50. To document variables, the object class, the property, and the value domain are specified, named, and defined. Each of these components can stand on its own and is reusable in the construction of other data elements. A general strategy can be adopted 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 statistical agencies.

51. All of the building blocks, and the links between them in the context of given statistical datasets or surveys, can be stored in a metadata repository. From there it is possible to produce, dynamically and on request, the complete definition of every variable, according to the specifications of the standard.

52. Consider the data element named "type of expenses of business location" as an example. It is analyzed as follows:

- "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. "Expenses" is the name of a property.
- "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. "Business location" is the name of an object class.
- "Type" refers to the reporting of "Expenses of Business Location" using the classification called Expense Categories, Annual Survey of Manufactures (ASM). "Type" is a representation term, naming the kind of value domain.

53. Most data element documentation can be approached in this straightforward manner but some situations arise where conventions or consistent approaches need to be developed and applied to deal with the varied ways used by data producers to present statistical data. Some practical choices need to be made in applying the standard and conventions adopted to deal with some of the more common situations. Suggestions for common applications are presented for each major construct in the standard: object class, property, and value domains.

ii. Object Class (Statistical Unit)

54. 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".[12] With reference to the standard, this makes it a statistically relevant type of object class, defined 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." In applying the standard, it is desirable to try to limit the number of object classes by using fundamental statistical units, if possible. Fundamental statistical units are defined as those that are not types of any other statistical unit and cannot be derived as grouping of any other statistical unit.[13] The following types of fundamental statistical units were identified:

- Agents: Entities that act and whose actions are reported on by statistical agencies. In social statistics, "Person" can be considered as an agent.
- Events: Actions of (or by) agents as reported by statistical agencies. Events are discrete in time (occur in a time period) and finite (can be counted). In social statistics "Birth" can be considered an event.
- Items: Things that are generally either produced or managed by agents. In economic statistics, "Product" can be considered as an item.

55. Fundamental statistical units are 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 should be 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.

56. Starting with fundamental statistical units and only identifying additional statistical units defined according to certain properties when these are commonly encountered, national statistical agencies can represent their data holdings with a fairly limited set of about 80 object classes. These are shown in the following table.

---

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

[13] 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.

Table 2 – Object classes defined in Statistics Canada's metadata registry

	Social	Economic
Agents	Family	Business Entity
	Child	Business Location
	Crime Victim	Earner
	Criminal Accused	Economy
	Criminal Charged	Employed Person
	Emigrant	Employee (LFS)
	Homicide victim	Employee – private
	Household	Employee – public
	Household Head	Employee (SEPH)
	Immigrant	Employment Insurance Beneficiary
	International Migrant	Enterprise
	Interprovincial Migrant	Establishment
	Intraprovincial Migrant	Farm Operation
	Mother	Institutional Unit
	Non-permanent Resident	Labour Force Participant
	Person	Paid Worker (labour income)
	Person 12 Years or over	Paid Worker (labour market)
	Person charged	Person 15 Years or over
	Smoker	Self-Employed Worker
	Student	Traveller
	Suspect - Chargeable	Unemployed Person
	Woman	
Events	Birth	Person-trip
	Community Admission	Person-visit
	Correctional Service Admission	
	Criminal Incident	
	Custodial Admission	
	Death	
	Divorce	
	Homicide	
	Marriage	
Things	Case	Building Permit
	Charge	Crop
	Criminal Offence	Employment Insurance Claim
	Dwelling	Farm Input
	Legal Aid Application	Help Wanted Ad
	Legal Aid Plan	Job
	Probation Order	Passenger-Kilometer
	Sentence	Product
	Shelter	Security
		Transaction
		Vehicle
		Vehicle –Kilometer
		Visit-night

57. Under the standard, almost anything can be an object class. In practice, certain choices must be made. A test that can be used in trying to isolate the object class is to answer the question “What is being counted here?” For example, we may be 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.

iii. Properties

58. Properties are simply the characteristics of interest of the unit of observation (object class). These include sex, income, industry type, and number of employees, among many others. Having defined the object class, it is relatively straightforward to identify which characteristics are being measured.

59. The application of the standard leads us to sometimes define “compound” properties. Limiting object classes to fundamental statistical units, only occasionally further qualified by a property, shifts more of the meaning to the property and value domain definitional space. As a result, it may be necessary in some cases to define a property with more than one dimension. For example, in the data element "number of production workers’ hours paid of business location", the property is "production workers’ hours paid". "Production workers" and "hours paid" can be considered object classes each on their own. They can be counted, they have similar characteristics, etc. However, in this dataset, the "business location" is the unit of observation and "hours paid" and "production workers" have been combined to form a compound property of the business location.

60. Another example where this is the case is the data element concept named "race of reference person of household", where "reference person" refers to the person in the household used to define all the relationships between members. Adding to the confusion, the US uses such data to apply "race" to the household in order to perform imputations for missing data. So, "reference person" has a subservient role in the semantics. Applying the question "what is being counted here?", the answer is not immediately clear. The object class could be "consumer unit" or "reference person of consumer unit", and the property contains the remaining semantics in each case: "race of reference person" or "race". The question reduces to whether the meaning refers to the data at data collection or during the analytical stage. Applying the list of available object classes (see table 2) solves the problem.

61. Based on the experience of Statistics Canada, it appears as if the specification, naming, and definition of about 500 properties will be sufficient to cover all statistical data published by a national statistical agency.

iv. Value domains

62. Value domains come in two types: enumerated and non-enumerated. They are the set of allowed values a data element may take. In ISO/IEC 11179, a value domain is defined as a set of permissible values, where a permissible value is a pair containing a value and its meaning. The set of value meanings is called a conceptual domain. Conceptual domains also come in the two types: enumerated and non-enumerated. A simple example of an enumerated value domain is "sex codes", which may contain the following permissible values: See ISO (2003) for a more complete discussion of value domains and conceptual domains.

<M, male>
<F, female>

63. The corresponding conceptual domain is the set of value meanings:
{male, female}

64. The enumerated and non-enumerated types correspond usually to the standard statistical datatypes of categorical and quantitative data, respectively. There are exceptions, but mostly these have to be manufactured. The natural way to use the types of value domains lends itself easily to the statistical datatypes.

65. In particular, 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). So, the value meaning for each value in a non-enumerated value domain is the unit of measure.

66. List of units of measures recorded in Statistics Canada's metadata registry are

Area in Acres
Area in Square Feet
Areas in Hectares
Basic Price in Current Dollars
Canadian Dollars
Chained 1997 Dollars
Constant Dollars
Count in Dozens
Count in Metric Bundles
Count in Metric Rolls
Count in pairs
Counts in Whole Numbers
Current Prices
Quantity in Megawatt hours
Set of Numbers Expressed as Indexes
Set of Numbers Expressed as Rates
Set of Numbers Expressed as Ratio

Time in Days
Time in Hours
Time in Years
Volume in Bushels
Volume in Gallons
Volume in Kilolitres
Volume in Litres
Volume in Quarts
Volume in Tonne-kilometres
Volumes in Cubic Metre-kilometres
Volumes in Cubic Metres
Volumes in Cubic Metres Dry
Weight in Hundredweights
Weight in Kilograms
Weight in Metric Tonnes
Weight in Pounds
Weight in US tons

67. 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 could be the (slightly out of date) population of Victoria, BC in Canada. As a code, it is actually 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: values (codes) and value meanings (labels), which can also have a definition.

68. 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 this application of the standard, statistical classifications are re-created from value domains specified according to the standard. As a "classification" entity does not exist in the standard, a number of conventions are therefore developed for this purpose.[14]

69. First, every value domain is 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.

70. In certain cases, classifications are altered by data producers by grouping certain classes together. This in effect introduces a new value domain, or a new level in a classification, that is not exhaustive of the universe of observations to be classified.

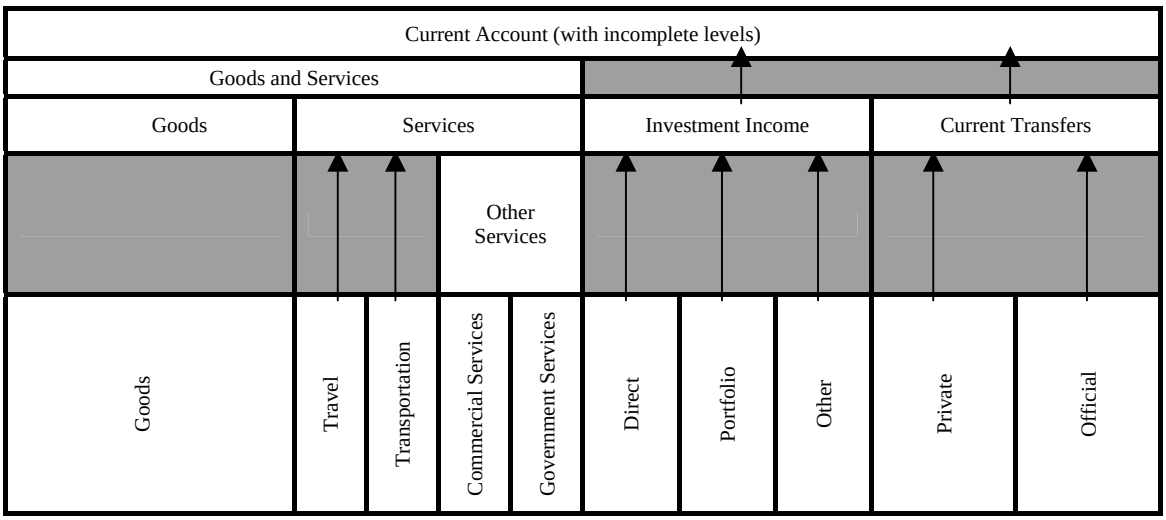
Table 3: A Value Domain for Current Account

Current Account									
Goods	Services				Investment Income			Current Transfers	
Goods	Travel	Transportation	Commercial Services	Government Services	Direct	Portfolio	Other	Private	Official

[14] 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.

71. Table 3 provides an example of a classification comprised of three value domains (levels), in this case the Current Accounts Classification[15] used in Canada. Table 4 shows how the classification has been altered by a data producer by grouping together "Goods and Services" on the second level and Commercial services and Government services on the third level.

Table 4: Alternative Value Domain for Current Account



72. This has in effect introduced value domains that are not exhaustive of the universe at levels 2 and 4 of the altered classification. If the classification is presented one level at a time to users, which could well happen in many applications, the user will have incomplete information concerning all the values that the data element being represented by this value domain could assume. To correct for this, every level is made to be exhaustive of the universe of observations to be classified by "promoting" classes from the level below (see arrows in Table 4). This results in a new 5-level classification, as shown in Table 5.

Table 5: A Corrected Alternative Value Domain for Current Account

Current Account (Incomplete levels filled in)								
Goods and Services				Investment Income			Current Transfers	
Goods	Services			Investment Income			Current Transfers	
Goods	Travel	Transportation	Other Services	Direct	Portfolio	Other	Private	Official
Goods	Travel	Transportation	Commercial Services / Government Services	Direct	Portfolio	Other	Private	Official

[15] This example is taken from Johanis, P., Brooks B., Dunstan T., and Lévesque, J-S. (2003), Statistics Canada's Implementation of the Data Element Model, paper for the Metadata Registries Open Forum 2003.

73. The outcome is a well configured classification, rectangular, exhaustive at every level, with classes that are mutually exclusive and exclusive in aggregation. Another advantage of this approach is that it preserves the relationship between the original classification and its variants. This promotes the reuse of standard value domains (in the example above, levels 1, 3, and 5 in the variant are identical to levels 1, 2, and 3 in the original) and clearly shows how one relates to the other.

74. 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.

75. 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.

North American Industry Classification System Canada 1997 (standard)
NAICS 1997 Durable / Non-Durable Manufacturing Industries
NAICS 1997 Energy Sector
NAICS 1997 GDP
NAICS 1997 GDP Finance, Insurance, and Real Estate
NAICS 1997 GDP Special Industry Aggregations
NAICS 1997 Goods and Services
NAICS 1997 ICT
NAICS 1997 ICT (Manufacturing/Services Split)
NAICS 1997 Industrial Production (GDP)
NAICS 1997 IOFD
NAICS 1997 Labour Income
NAICS 1997 LFS
NAICS 1997 Trade Groups

76. 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 the 2002 version were considered new versions of these variants. 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 are containers of value meanings, which are re-used in value domains.

77. This application of ISO/IEC 11179 to document statistical classifications is consistent with the approach developed by the Neuchâtel group (see section V.B) but has not been reconciled in detail. Uncovering the meta-models underlying the various "classification servers" in use around the world and developing a common approach to documenting

classifications, based on this approach or another, would be a major, and attainable, achievement for the world statistical community.

v. Naming

78. For every item we care about the meaning, we give it a name. Names are the means by which people remember things and first infer some meaning. The ISO/IEC 11179 devotes an entire part of the standard to naming: Part 5.

79. The naming convention for data elements provided in the standard is quite simple. Data elements names consist of main parts:

- Object class term
- Property term
- Representation term

80. The first two of these are understandable from the previous sections. The representation term refers to the value domain and other attributes associated with the representation of the data element. Here "representation" includes the allowed values, their datatype, and the unit of measure (if necessary). Representation is the "form" of the data as they appear on paper or the screen.

81. Terms such as count, value, and number are used as representation types in the case of data elements with non-enumerated value domains. "Value of expenses of business location" is an example of such a data element name, 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 name "category of age of person". In the end, a relatively small set of representation types is sufficient to cover all of the statistical output of a statistical agency (see list below).

Table 6 – Representation classes defined in Statistics Canada’s metadata registry

Enumerated	Non-enumerated
Category	Amount
Code	Area
Level	Average
Name	Duration
Status	Index
Type	Length
	Mean
	Number
	Percentage
	Proportion
	Quantity
	Range
	Rate
	Ratio
	Value
	Volume
	Weight

vi. Data Elements

82. With these elementary building blocks defined, it is possible through combinations of object classes, properties, and value domains to specify, name, and define all data elements produced by a national statistical agency. For Statistics Canada, this approach has resulted in the identification of around 900 data elements, covering the entirety of its statistical output disseminated through CANSIM.[16] This is where the consistent application of ISO/IEC 11179 could yield major harmonization gains across national statistical agencies.

83. 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.

84. The problem of interoperability, unfortunately, is not that simple. There are considerable disharmonies in the value domains, or classifications, used to represent these data elements. So, work to harmonize classifications (and value domains in general) across statistical agencies is required to make data interoperable.

IV. CORPORATE METADATA REPOSITORY (CMR) MODEL

85. This section contains a partial description of the CMR model. Each of the sub-sections contains a model represented in the Unified Modeling Language (UML). These models are less detailed versions of those found in the CMR model. The attributes depicted in each class are meant to signify greater detail in the actual model.

86. The model presented is a conceptual model. Datatypes are not provided with the attributes, and there are not enough details to insure consistent implementations. Nevertheless, the US Census Bureau, Bureau of Labor Statistics, and Statistics Canada each have implementations of this model.

A. Data Dimension

87. The data dimension describes the semantics of data. It is a copy of part the model specified in ISO/IEC 11179. All the provisions needed for consistent implementation are described in the standard. If one were to implement the standard as part of a CMR implementation, the result would be a conforming[17] implementation of ISO/IEC 11179 (ISO, 2005).

88. The following diagram is a high level overview of the data model. The classes in the model are described in section III.

[16] This understates the actual count slightly as there are a few data elements for which metadata are still in a staging area and have not been loaded into the IMDB.

[17] The term conforming is defined in ISO.IEC 11179-3: Part 3 - *Metamodel and basic attributes*.

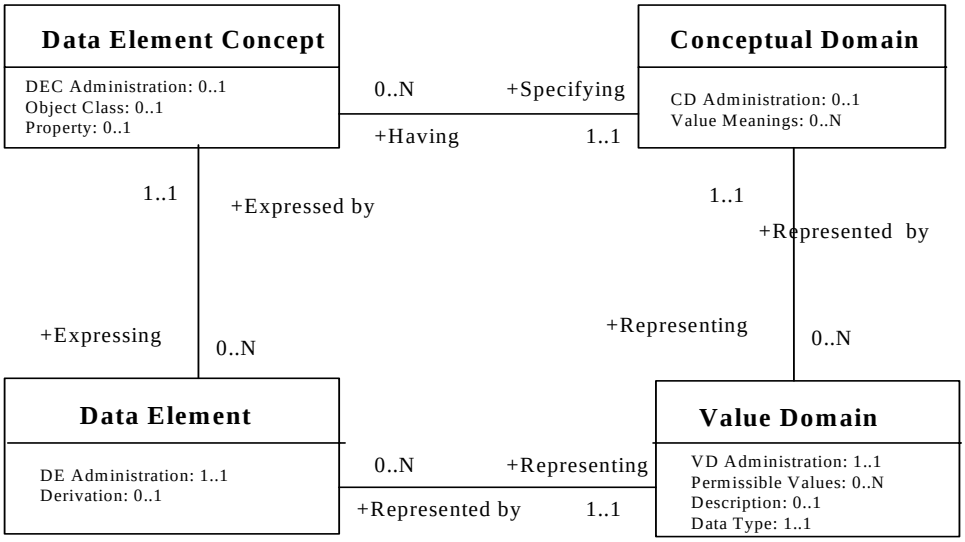


Figure 2: Data Dimension Model

B. Business Dimension

89. The Business Dimension describes the business of statistical organizations. It is composed of classes, attributes, and relationships that describe data that the organization needs to keep about surveys. The model supports the storage of metadata as single attributes or as documents. Figure 3 shows the Business Dimension model.

90. The model describes survey designs, processing, analyses, and data sets. It contains classes for each of the important parts of a survey. The model supports organized storage and complex searches for metadata describing a survey, and it supports searches for metadata across multiple surveys. The model also provides several other features:

- A list of all current surveys conducted by the agency
- Comparison of designs, specifications, or procedures across surveys
- Reuse of designs, specifications, or procedures
- Categorizing and classifying documents
- Assembling complete documentation for a survey
- Attributes to support embedded metadata

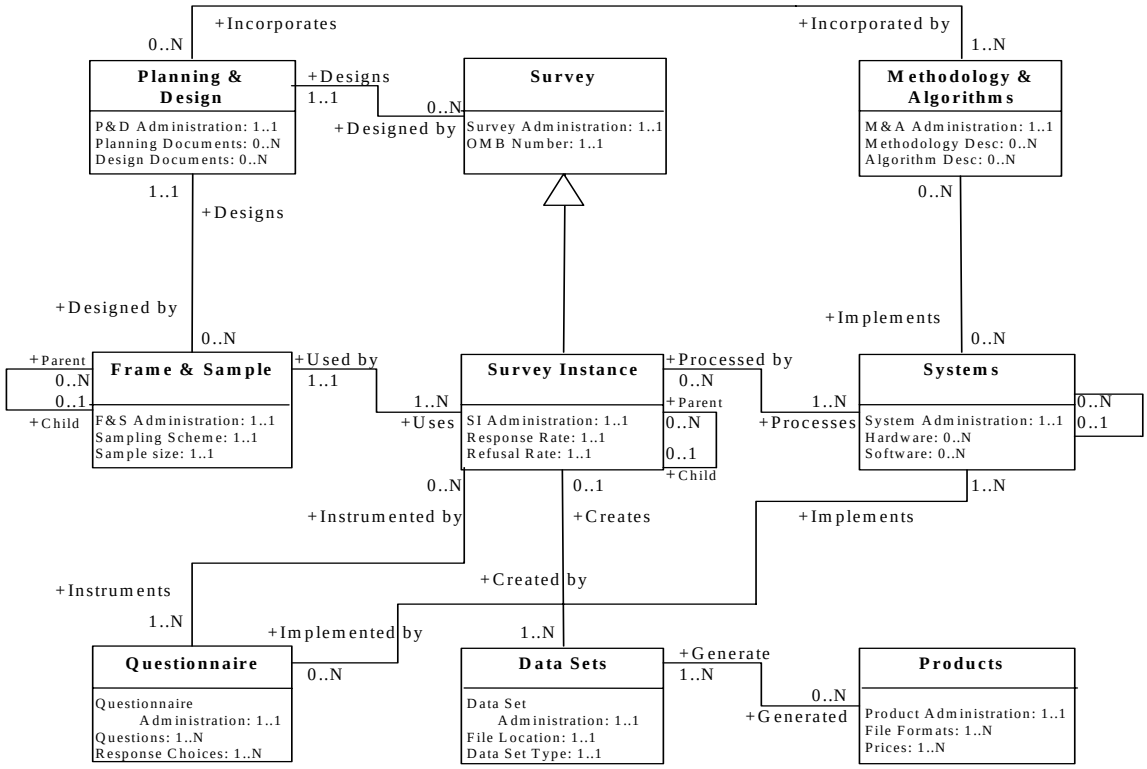


Figure 3: Business Dimension Model

i. Survey Life Cycle

91. The model supports the survey life cycle. Content, Planning, and Design are captured in the Planning and Design, Frame and Sample, Methodology and Algorithms, and Survey classes. Collection is captured in the Questionnaire, Survey, Survey Instance, Data Set, and Systems classes. Processing and Analysis are captured in the Survey Instance, Methodology and Algorithms, Systems, and Data Set classes. Finally, Dissemination is captured in the Data Set, Systems, and Product classes.

ii. Questionnaire Model, Linking Business – Data Dimensions

92. The CMR model contains many links between the various dimensions within the model. Using the detailed questionnaire model, Figure 4 illustrates links between questionnaires and data elements. Data Element Concepts and Questions are linked, because they each express concepts describing the same data, albeit from a different perspective. Value Domains and Response Choices each describe the valid values some data can take.

C. Administration and Document Dimensions

93. The Registration Authority (ISO, 2005) establishes the rules under which the repository operates. Monitoring metadata quality, monitoring the life cycle of the described objects, and maintaining paths of accountability for metadata are important functions.

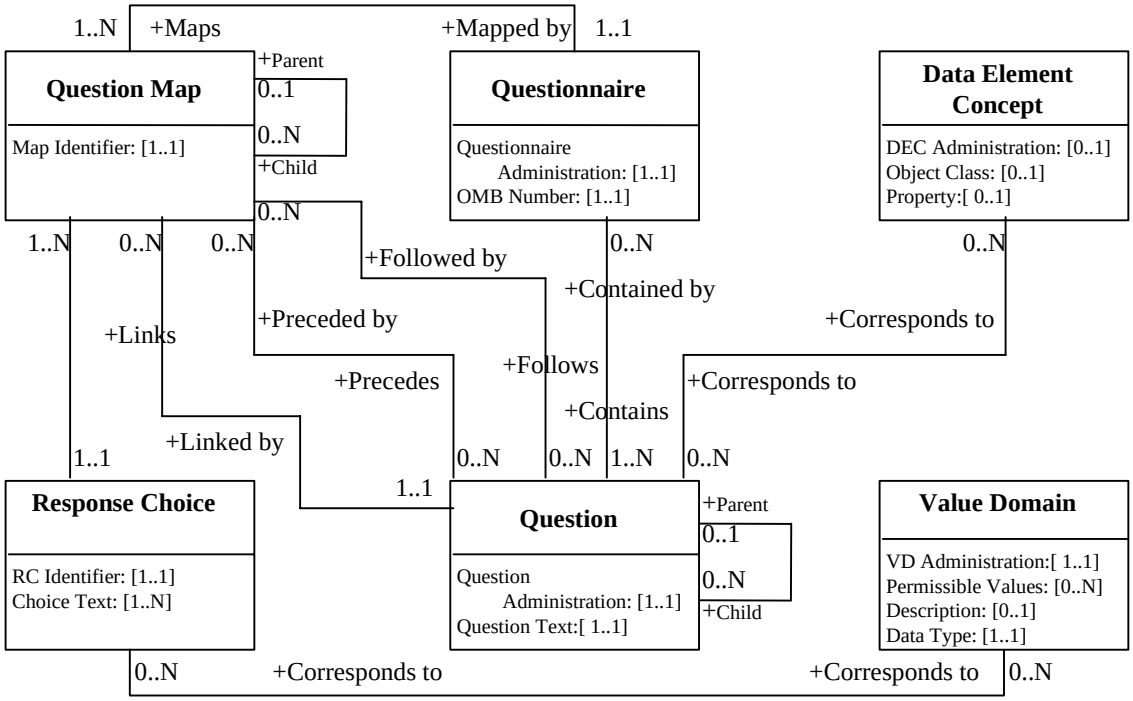


Figure 4: Questionnaire Model

94. An Administration Record is established each time an object is described, or *registered*. The common attributes are provided along with specialized attributes for each object. Metadata is often provided in the form of documents, so URL's to relevant documents are critical metadata. The model allows links to as many documents as necessary. Each document may be linked to many objects. Figure 5 provides a data model for the registration process.

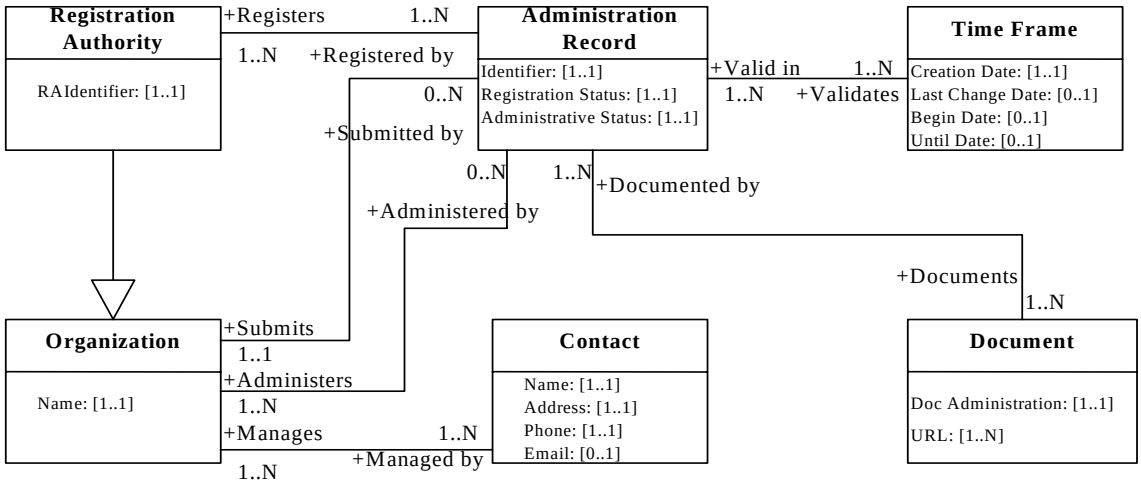


Figure 5: Administration and Documents Dimensions Model

D. Terminology and Classification Dimensions Model

95. The CMR model contains a dimension for managing classification schemes used to classify objects the CMR model describes. The term "classification scheme" is slightly

misleading. The model supports representing a concept system. Classification is achieved by relating objects (administration records) to other concepts. The classification of objects ties them to particular subject fields, such as the concept system determined by the variables of interest for a survey. In other words, the concepts a survey is trying to measure make up part of the subject field for that survey. For instance, assigning an object class and property, as described in section III.C, is a form of classification.

96. The CMR model also supports terminology management as it applies to the concepts, terms, and characteristics that describe registered objects. One of the most important metadata elements for any registered object is its definition. This is crucial for understanding the meaning of the object. The CMR model supports *semantics*. Terminology management is a fundamental part that aim. Figure 6 provides the terminology and classifications model.

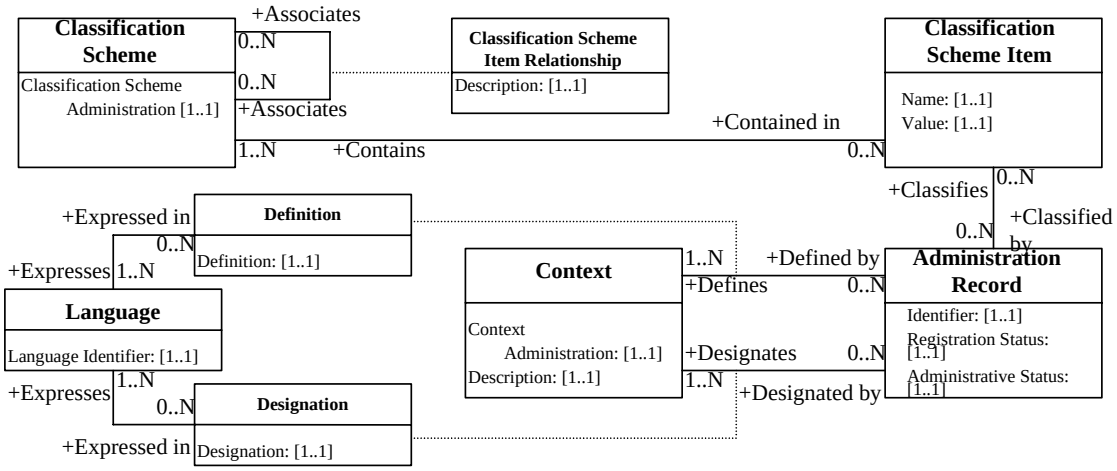


Figure 6: Terminology and Classification Dimensions Model

V. METADATA SCHEMES

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

A. eXtensible Business Reporting Language (XBRL)

98. XBRL (eXtensible Business Reporting Language) is an XML-based, royalty-free, open standard for business reporting. XBRL was developed by XBRL International, a not-for-profit consortium of over 400 leading companies and organisations around the world. Its main focus is to provide a standard format for recording financial data and associated metadata for financial statements and other business data reports. As these are frequently collected by statistical agencies as part of their economic statistics programme, XBRL provides a useful, and perhaps eventually universal, format for electronic data reporting for businesses. In this event, it would be very desirable if the metadata associated with the reported business data could automatically be converted from the XBRL format to the format used in the data repositories of statistical agencies.

99. In XBRL, the main organizing concept is the “instance document”, which represents one instance of some kind of financial report. In a statistical application, it is equivalent to a completed questionnaire or collection instrument. It therefore fits the CMR questionnaire model as described in figure 4.

100. Each instance document refers to a taxonomy, which describes the elements used in the document (for example, fixedAssets, totalAssets, subscribedCapital, and totalLiabilities), as well as the relationships among the elements. XBRL provides a structure to record the metadata for the concepts to be reported and cross-concept relationships, expressed according to the rules of XML syntax in the form of linkbases.

101. An XBRL taxonomy contains the following:
- **Schema:** A group of structured elements that may be used in instance documents. It is a dictionary of defined terms.
- **Label linkbase:** Labels or text associated with the elements in the dictionary may be created in different languages and used for different purposes.
- **Reference linkbase:** References to legal texts or accounting standards on which the concept is based.
- **Presentation linkbase:** Rules to specify the hierarchical relationships between elements.
- **Calculation linkbase:** Rules for calculations (additions and subtractions) between elements in the taxonomy.
- **Definition linkbase:** Rules to document other types of relationships between elements in the taxonomy.[18]

102. The challenge is to extract and re-interpret the metadata available in an XBRL taxonomy, that is the schema and all the referenced linkbases, according to the metamodel of ISO/IEC 11179/CMR. Fundamentally, this means identifying all the data elements and associated value domains that are embedded in a taxonomy.

103. The approach for doing this follows the procedure described in section III.G of this paper. What is the object class? For business reporting, it is always the enterprise. What are the properties of an enterprise being measured? First, there are identification properties such as geographic location, and industry. These can be defined in a straightforward way as per section III.G, i.e. name of geographic location of enterprise, type of industry of enterprise, etc. For the properties that describe the financial position and performance of the enterprise, however, we have different choices, depending on how economical we wish to be in identifying data elements. Every element in an XBRL taxonomy could be defined as a property, so we might define data elements such as value of fixed assets of enterprise, value of financial assets of enterprise, value of intangible assets of enterprise, value of total assets of enterprise. This would yield a very large number of data elements, many of which would by definition be interdependent (for example, total assets must be the sum of fixed, financial and intangible assets). An alternative is to consider each such element to be a permissible value within a value domain associated with a more aggregated data element, such as Type of asset of enterprise. This would yield two data elements, Type of asset of enterprise and Value of asset of enterprise. The former would have an enumerated value domain (with three values: fixed, financial, and intangible), the later would have an unremunerated value domain.

104. With this approach in mind, it is possible to consider the mapping between the metadata structure within an XBRL taxonomy and a metadata standard such as ISO/IEC 11179/CMR.

B. Neuchâtel Group

105. The Neuchâtel Group is currently composed of representatives from Statistics Netherlands, Statistics Norway, Statistics Sweden, Swiss Federal Statistical Office, and the

[18] This section draws heavily on a White Paper on XBRL Concepts and Recommendations, published by the Technology Working Group of XBRL Spain, September 2005; available on xbrl.org

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.

106. 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.

107. 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.

108. 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.

109. 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.

C. DDI

110. 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.

111. 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.

112. 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

113. 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.

114. 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 five schemes that is specifically engineered for describing statistical surveys. The Neuchâtel Group, described above, also produces standards directly related to statistical offices, but they are much more specialized.

115. 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.

116. The relationship between the ISO/IEC 11179/CMR standard and DDI has been frequently addressed in previous papers.[19] More recently, the feasibility of automatically generating DDI compliant metadata sets from an ISO/IEC 11179/CMR repository has been examined, which concluded that it was possible to generate almost all the metadata required by the DDI standard in this way.[20]

D. Statistical Data and Metadata Exchange (SDMX)

117. 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.

118. 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.

119. The SDMX model is very sophisticated and general and, while it has been developed for the exchange of statistical data and metadata, the generality of the model could satisfy the needs of many kinds of businesses to transfer many kinds of data and metadata.

120. SDMX provides a metamodel for documenting and structuring data sets and associated metadatasets. Provision is made for datasets in the form of time series, cross-sectional tables or data cubes. From the data perspective, the main construct in SDMX is the *data structure definition* and related classes. Many of the components of data are described here. A *data structure definition* in SDMX corresponds to a dataset in the ISO/IEC 11179/CMR approach. Related attributes such as concepts and code lists in SDMX can be related to constructs in ISO/IEC 11179 such as data element concept, data elements and value domains. The detailed mapping between the two models has not been checked in a rigorous way, but there is a strong relationship between the standards.[21]

---

[19] See Bradley, W.J., Colquhoun, G. and Ryssevik, J., **Integrating ISO/IEC 11179 and the DDI in a Single Application**, Presentation to Open Forum on Metadata Registries, Statistics Section, Santa Fe, January 2003; Bradley, W.J., Gillman, D.W., Johanis, P. and Ryssevik, J. **Is your Agency ISO Compliant? Standardizing Metadata for Improved Knowledge Delivery in National Information Systems**, Bulletin of the International Statistical Institute 54th Session Proceedings, Berlin, August 2003; Ryssevik, J., Glover T. and Colquhoun, G., **Relationship to ISO/IEC 11179,** Data Systems and Standards Division, Health Canada, 2003

[20] **Discovering Microdata Variables: Comparing DDI compliant documentation to an ISO/IEC 11179 metadata registry** Tim Dunstan, Statistics Canada, Chuck Humphrey, University of Alberta, Canada, Paper presented at Symposium 2005, Statistics Canada, October 2005.

[21] For a first attempt at such a mapping, see SDMX, ISO 11179 and the CMR, Arofan Gregory and Chris Nelson, METIS 2006

121. 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.

E. Common Warehouse Metamodel (CWM)

122. The Common Warehouse Metamodel (CWM)[22] is a standard developed under the auspices of the Object Management Group (OMG). Version 1.0 was published in February 2001. The CWM is among an integrated family of standards including Unified Modeling Language (UML, ISO/IEC 19501), Meta-Object Facility (MOF, ISO/IEC 19502), and XML Metadata Interchange (XMI, ISO/IEC 19503).

123. The CWM is a framework for integrating software and tools in the "information supply chain" (ISC). The ISC is a generalization of the survey production life-cycle in statitistical offices, where the typical stages are conception, design, collection, processing, analysis, and dissemination.

124. In the ISC, parts of the business data process are linked together by data flows as data move between systems. Metadata are managed by each system independently, in proprietary ways, without regard to any interoperability. The question of how to make metadata sharable throughout the ISC is the purpose of the CWM.

125. The CWM contains an extensible model (a metamodel, in the terminology for OMG)) for modeling metadata. The common modeling framework allows then for mapping between metadata models. One then maps metadata across the ISC, using the metadata output of one process to help drive the next.

126. The modeling facility has 5 layers, called Object, Foundation, Resource, Analysis, and Management. The Object layer contains core modeling elements used to construct metadata models, all based on a special subset of UML. The Foundation layer defines services such as datatyping, business information, and key indexing to provide the means to define modeling constructs (e.g., datatypes, business mappings, and keys) needed to model any kinds of metadata. The Resource layer defines the kinds of data resources in the ISC, for instance relational, multidimensional, or XML. This allows descriptions, say, of relational databases. The Analysis layer provides additional models used to support analysis tools, such as data mining and information visualization. Finally, the Management layer provides for processes and operations that one might use to interact with a data warehouse.

127. The major benefit to the CWM is the possibility for a statistical office to tie together all the systems used for a survey or groups of surveys into one integrated system. Normally, metadata interchange is done between pairs of systems, and a mapping must be built for every pair. The CWM provides for a single point of communication, so each component of a system must map its metadata to a CWM implementation. The CWM then provides for all the mappings between metadata schemes. Therefore, the complexity of an architecture based on CWM is much simpler. There are fewer metadata mappings that must be made.

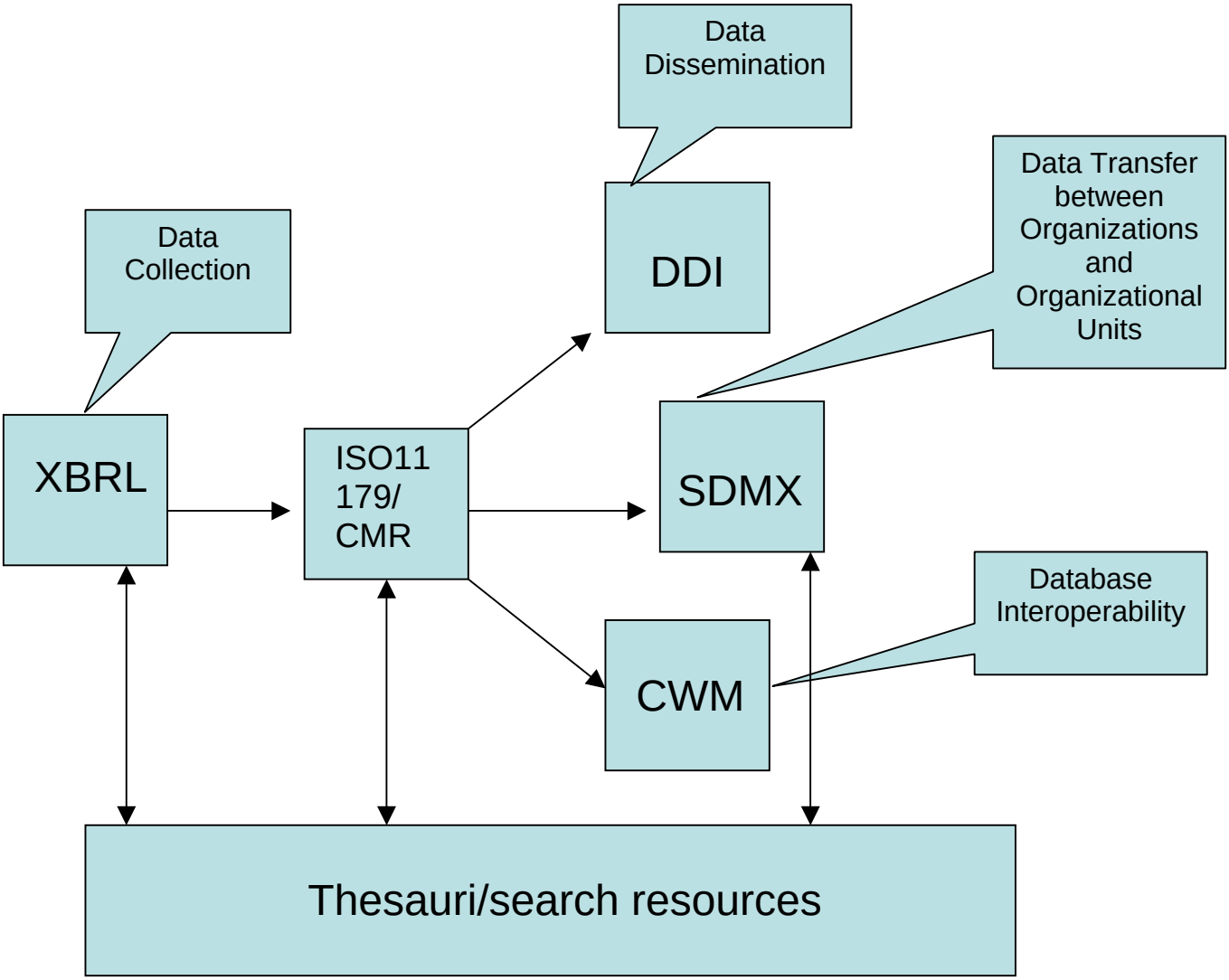
[22] OMG. (2000). *The Common Warehouse Metamodel*. Object Management Group

VI. Integration

128. There are two ways to consider integrating the standards described in section V in a statistical office. They are, loosely speaking, a systems integration and a conceptual integration.

129. From the point of view of conceptual integration, the standards can be related as shown in Figure 7. The semantics, or meaning associated with data elements, should be contained in the electronic data collection instruments according to standards such as XBRL and other EDR metadata standards. It should be possible then to map the names, definitions and other metadata attributes from this collection metadata to the statistical offices central metadata repository, structured according to the ISO/IEC 11179/CMR model, or some other comparable standard. From this repository, metadata sets should be generated automatically according to output standards such as DDI, in the case of microdata sets, SDMX in the case of transfers and exchanges of aggregated data sets between statistical offices, and CWM when statistical data are made available for analysis in data warehousing environments. It should also be possible to automatically generate other metadata sets structured according to agency specific needs, such as the SDDS standard of the IMF, or a statistical agency's own data quality statement standard.

Figure 7: Conceptual integration of Metadata standards for Statistical Office



130. A deeper integration can be conceptualized if systems integration is also taken into account. The CWM is the main hub for the systems integration, and the CMR (including ISO/IEC 11179 and the Neuchâtel models) for the conceptual part. This division is not absolute, as the CWM has a semantic component and the CMR may be useful for driving systems. Here, we use each to their strengths. See Figure 8 for a schematic diagram representing these ideas.

131. As described in section V.A, the CWM is used to integrate disparate tools by providing a common metadata framework for describing the metadata model for each tool. The CWM provides the functionality to map between metadata models, and in this way, the tools may "talk" to each other. By this we mean that metadata from one tool is used by another.

132. The situation of a set of tools talking to each other throughout the survey life-cycle was described by Kent, et al (1998). There, they described an automatic survey processing paradigm driven completely by metadata. The need for each tool in the survey life-cycle to understand the metadata model of the tool before it is of primary importance. The CWM provides a solution to this problem - see Figure 8.

133. On the other hand, there is still the need for understanding everything within the survey-life cycle. By this we mean the needs of humans to comprehend the systems they build, maintain, integrate, and upgrade. This does not necessarily mean all the systems are computerized, however, the ultimate aim is to automate as much as possible.

134. The CMR, including ISO/IEC 11179, and the more detailed models for statistical agencies developed by the Neuchâtel Group deal primarily with semantics. As described in section III, the ISO/IEC 11179 standard is a realization of the terminological approach to data. This approach, as described, is concept based. The Neuchâtel Group model for statistical data follows the ISO/IEC 11179 approach but extends the concepts to also describe data sets, cubes, and tables. The classification model from the Neuchâtel Group thoroughly describes classifications, mappings between them, and their management. Along with the extensions of the CMR to describe questionnaires and samples, a complete conceptual description of a survey possible.

135. Figure 8 in this section pictorially represents what we have discussed here. There are six parts to the survey life-cycle depicted, and they are linked serially from the conception to dissemination. The links represent the transfer of data or metadata as needed between the parts. Each part includes the systems and tools in place to complete that par of the life-cycle. There is no requirement that these be automated. In any case, there is metadata in use for each part of the life-cycle and it is represented by the database symbol. Then, there is a link between the local metadata model and the CWM, which describes and maps the models.

136. In addition, there are three other standards mentioned, for input and output of statistical data: XBRL for reporting from respondents; DDI for transferring metadata to users of statistical data; and SDMX for transfer of data and metadata to other statistical offices. These standards have links back to the CWM because they are metadata models themselves. Finally, there are links to the CMR and Concepts & Terms databases because they too are metadata models. This way the CWM contains a map for all the metadata models the office uses.

137. The CMR is used for describing statistical data and the survey life-cycle. The purpose is to provide descriptions at the conceptual level, creating a concept system for the statistical office; although attributes for driving systems are contained there, too. The cloud

emanating from the CMR database icon in Figure 8 is meant to convey this conceptual purpose.

138. The special terms and concepts used in a survey form a special language, and these concepts are the ones that appear as part of the descriptions in CMR instances. As described earlier, the object classes, properties, conceptual domains, value meanings, and data element concepts used to describe data are concepts in these special languages. Survey concepts, such as universes, populations, characteristics, and those described in questions (in a questionnaire) are also concepts in survey special languages. Therefore, we need to manage concepts and the terms designating them.

139. So, the CMR takes the concept system of the statistical office and gives it purpose. The purpose is the description of surveys and their data.

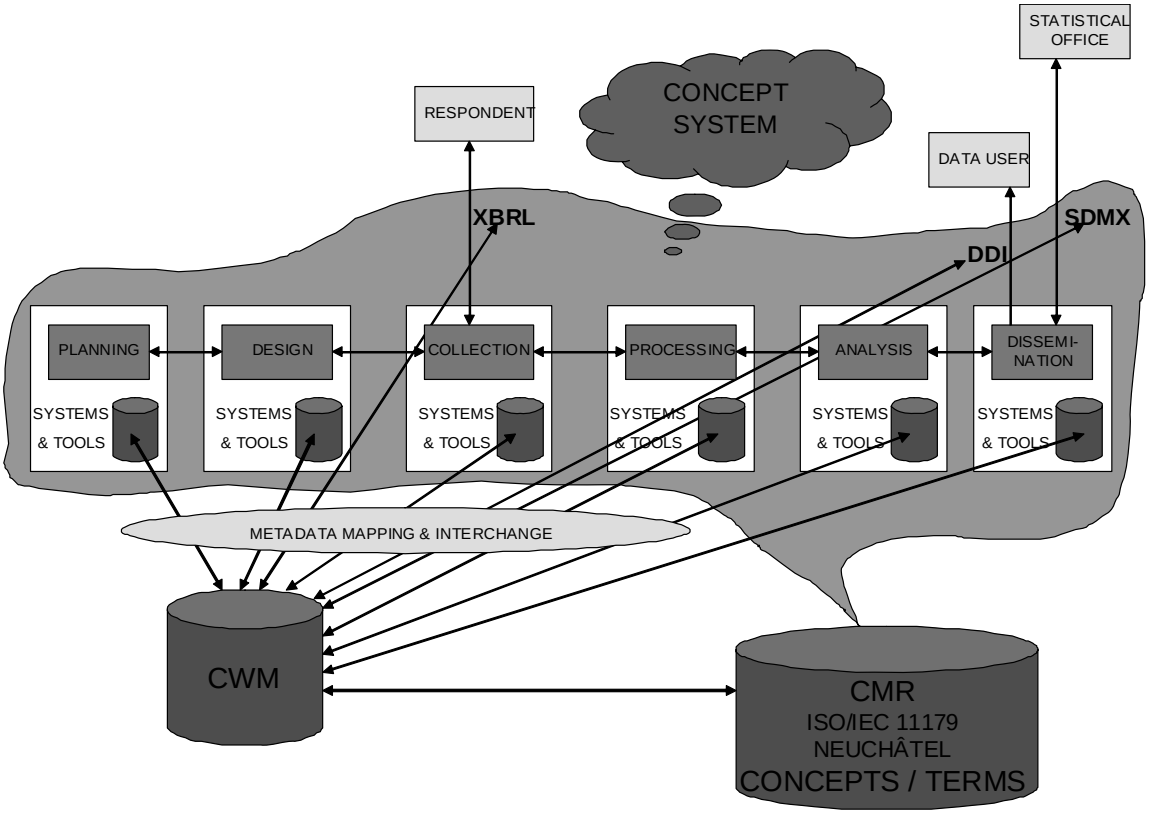


Figure 8: Metadata Architecture for Statistical Office

VII. CONCLUSION

140. Over the last few years, metadata systems and approaches for statistical offices have stabilized and matured. There is now a general recognition of what we mean by metadata and its form and function in statistical offices. A certain ecology of sometimes competing, sometimes complementary standards has also emerged and, to a certain extent, the outline of a process of natural selection can now be distinguished. This paper attempts to present the most prominent current members of this family of metadata standards and how they can be articulated to serve the needs of statistical offices. It should now become apparent that the

goal is not to develop a single overarching standard that covers everything. Rather, a limited set of well developed standards, appropriately adapted to specific statistical junctions, can and should become the target for world adoption and interoperability. A number of specific and achievable convergence targets have been identified in this paper and more could be developed by the METIS Task Force, in the context of an integrated framework such as the one presented in this paper. Given this advanced stage of development, continued sharing of experiences and expertise, but focused on consolidation and harmonization of approaches, should represent the next phase of the international collaboration undertaken under the METIS work program.

VIII. REFERENCES

Bradley, W.J., Colquhoun, G. and Ryssevik, J., *Integrating ISO/IEC 11179 and the DDI in a Single Application;*

Bradley, W.J., Gillman, D.W., Johanis, P. and Ryssevik, J. *Is your Agency ISO Compliant? Standardizing Metadata for Improved Knowledge Delivery in National Information System*, ISI Conference 2004, Berlin

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.

Dunnet, G. and Merrington, R. (2006). *Statistics New Zealand's end-to-end metadata life-cycle*. Working Paper #16 presented at the UNECE Workshop on Statistical Metadata. Geneva, Switzerland.

Dunstan, T., Humphrey, C., *Discovering Microdata Variables: Comparing DDI Compliant Documentation to an ISO/IEC 11179 Metadata Registry*, Symposium 2005, October 2005, Statistics Canada

Farance, F. & Gillman, D. (2006). Working Paper #12 presented at the UNECE Workshop on Statistical Metadata. Geneva, Switzerland.

Froeschl, K., Grossmann, W., & 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. (2006) Theory and Management of Data Semantics. In D. Schwartz (ed.) Encyclopedia of Knowledge Management. Hershey, PA, USA: Idea Group.

Gregory, A., Nelson, C., *SDMX, ISO/IEC 11179 and the CMR*, METIS 2006, Geneva

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. (2003). *ISO/IEC TR 20943-3: Procedures for achieving metadata registry content consistency, Part 3: Value domains*

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

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

Johanis, P., Brooks B., Dunstan T., and Lévesque, J-S. (2003), *Statistics Canada's Implementation of the Data Element Model*, Open Forum, Santa Fe

Kent, J.-P. (1998). *Take Care of the Meta, and the Meta Will Take Care of the Data*. Working Paper #6 presented at the UNECE Workshop on Statistical Metadata. Geneva, Switzerland.

Kutin, J. and Arnic, A. (2004). *Recent development of SORS metadata repositories for a faster and more transparent production process*. Working Paper #21 presented at the UNECE Workshop on Statistical Metadata. Geneva, Switzerland.

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

Mechanda, K., Johanis, P., and Webber M., (2003) *Conceptual Model for the Definitional Metadata of a Statistical Agency*, Open Forum, Santa Fe

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

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.

OMG (2000). *The Common Warehouse Metamodel*. Object Management Group

Poole, J., *et al* (2002). *The Common Warehouse Metamodel*. New York: John Wiley and Sons.

Ryssevik, J., Glover T. and Colquhoun, G., *Relationship to ISO/IEC 11179*

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

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

Sundgren, B. (1995). *Guidelines for the Modelling of Statistical Data and Metadata*. Conference of European Statisticians Methodological Material, United Nations, Geneva.

Technology Working Group of XBRL Spain, *White Paper on XBRL Concepts and Recommendations*

W3C (2004). *Extensible Markup Language*. XML 1.1 reference specification

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