WP. 8
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)

Topic (ii): Metadata Concepts, Standards, Models and Registries

SDMX, ISO 11179 AND THE CMR

Invited Paper

Submitted by SDMX Standards Team[1]

---

[1] Prepared by Arofan Gregory and Chris Nelson.

SDMX, ISO 11179 and the CMR

Arofan Gregory

Chris Nelson

Metadata Technology Ltd. April 2006

Contents

1 Purpose and Scope of the Paper

1.1 Purpose

ISO/IEC 11179 is concerned with the semantics of data, whilst SDMX is concerned with the structure of data. In order to achieve their respective goals, the two standards have some similar constructs and, whilst these could be compared, this is not a very productive exercise. The intent of the standards is different and the way each use these constructs is different.

Therefore it is not the intent of this paper to explore whether ISO 11179 semantics can be described in SDMX, or whether SDMX structures can be described by ISO 11179, as the answer is “perhaps, but this is not how the standards should be used”. The intent is to explore whether and how ISO 11179 Data Elements can be constructed from SDMX data and metadata structures, or instances of these structures, so that organisations can map ISO 11179 repository items with SDMX repository items.

It is also the intent of this paper to describe briefly the way the Corporate Metadata Repository model (CMR) uses ISO 11179 to capture the underlying semantic of key structures in the CMR, and to describe the scope of the registry models in ISO 11179 and SDMX, to show how these can interact.

Note that, unless referenced in a specific context, ISO/IEC 11179 is referred to hereafter as ISO 11179.

1.2 Scope

The scope of this paper is summarised in the diagram below.

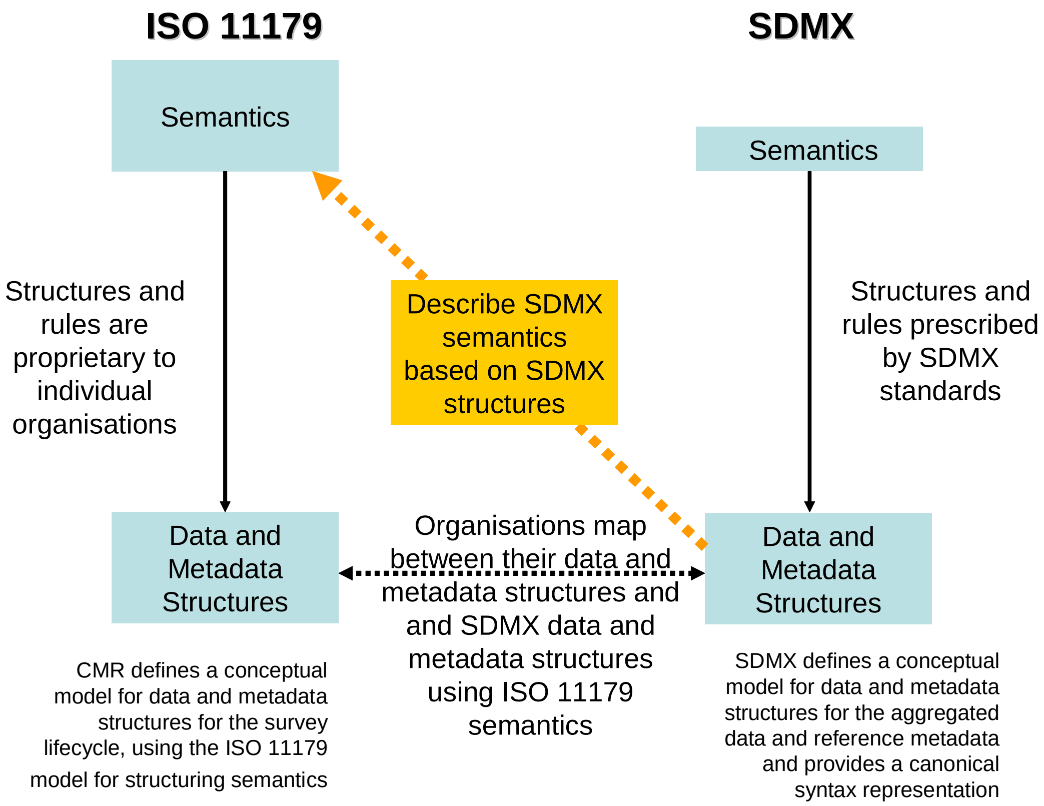


Figure 1: Schematic of the possible scope of using ISO 11179 to describe the semantics of SDMX constructs

The scope of this paper is show how data elements semantics can be derived from SDMX structural definitions, and how and why this mapping differs from the way this is achieved in CMR.

This paper takes as its base Part 3 (Registry metamodel and basic attributes) and Part 5 (Naming and identification principles for data elements), of ISO/IEC 11179 and provides a mechanism for naming data elements based on the SDMX Data and Metadata Structure Definitions.

2 Scope of the Standards

2.1 ISO 11179

"Some standards describe content and not format, metadata standards describe the data necessary to describe other data or processes, and data semantics refers to the meaning of data. ISO/IEC 11179 is a metadata content standard focused on the semantics of data".

(Daniel Gillman "CORPORATE METADATA REPOSITORY (CMR) MODEL" EU Metanet R&D Project Stockholm 2002)

ISO 11179 is not concerned with the structure of data: the metamodel is concerned with the artefacts required to describe the semantics of data. There is no canonical representation of ISO 11179 in syntactic form such as an XML schema.

2.2 CMR

The CMR extends the ISO/IEC 11179 (Part 3) metamodel in the field of statistics. Specifically the CMR model is designed to support:

- metadata necessary to describe the survey life cycle
- linkages between similar designs and processes used across surveys
- use of metadata to drive systems in support of the survey life cycle

The CMR model is a conceptual model and like ISO 11179 there is no canonical representation in syntactic form.

2.3 SDMX

The SDMX metamodel is concerned very much with the structure of data and metadata and with the semantics required to understand the meaning of the data and metadata carried in these structures. It is not concerned per se with modelling structures for defining the semantics of data elements.

Therefore, it should be possible to define an SDMX object (i.e. an instance of an SDMX class such as a Data Set or Observation) in terms of the semantics defined in the ISO 11179 metadamodel. How this can be achieved is detailed below.

SDMX is also concerned with syntax specifications, and harmonising content standards such as the names of "Concepts".

The SDMX specifications encompass:

- Information Model
- Syntax implementations based on the Information model (XML schemas and UN/EDIFACT)
- Content

- Statistical subject matter domain scheme
- Cross-domain concepts
- Metadata terminology (MCV – Metadata Common Vocabulary)

The diagram below shows the way these three key deliverables interact:

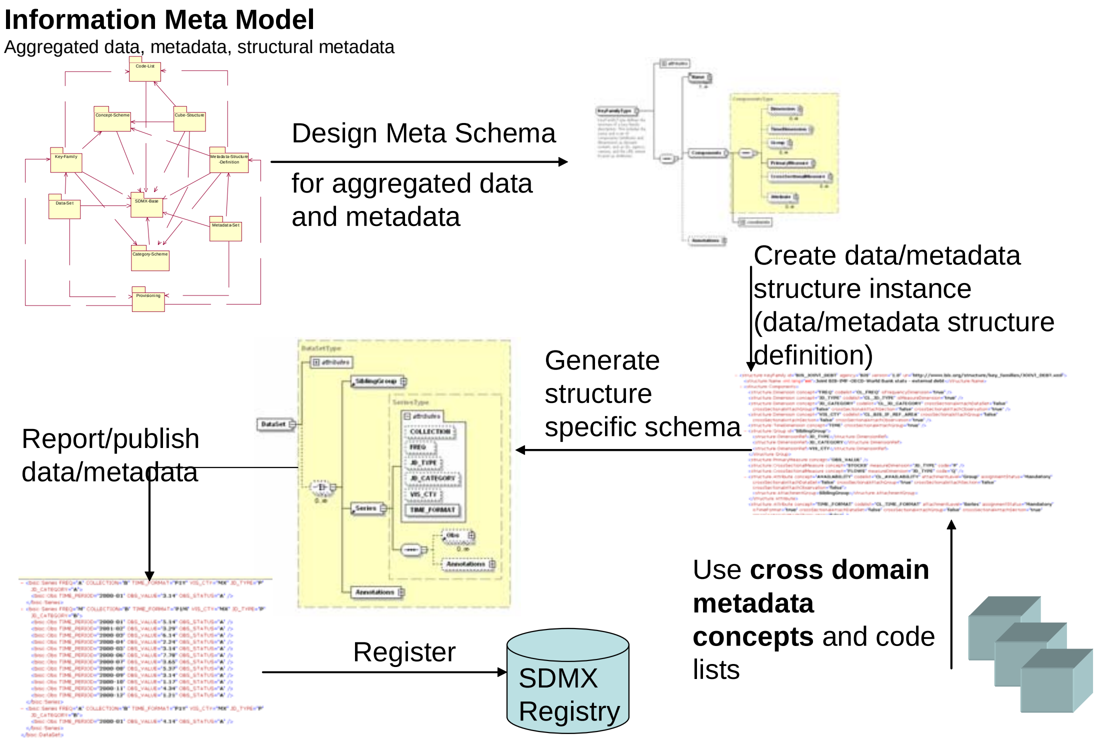


Figure 2: Interaction of key SDMX deliverables to create data and metadata schemas

The Information Model defines the artefacts necessary to support data and metadata publishing and exchange, and the way these artefacts are associated with each other. The model is used as the source of the syntax specifications – there are specifications for both XML representations (SDMX-ML) and UN/EDIFACT representation (SDMX-EDI) of the model, though the SDMX-EDI is limited to supporting just the data side and the necessary structural metadata to describe this. The diagram above depicts the XML scenario. The meta schema specifies the way to define a Data Structure Definition (also known as a Key Family), or a Metadata Structure Definition. A specific Data or Metadata Structure Definition (e.g. for External Debt, or Demography statistics) is defined in an instance of this meta schema. In order to achieve interoperability at the semantic level as well as the syntactic level, SDMX has worked on content standards and in particular on the definition of cross-domain concepts. Data and metadata structure designers are encouraged to use these concepts wherever possible, including the "core" representations (i.e. allowed format and code lists) defined for them.

The individual Data and Metadata Structure Definition is then used as the basis for generating a specific XML schema that supports the publishing or exchange of a data set or metadata set for that specific type of data or metadata. Therefore, a schema for reporting Balance of Payments data would be different from the schema generated to report Labor statistics. However, the difference is only in the naming of the XML tags, which reflect the names of the concepts used in the structure definition. SDMX defines a standard way of generating these

schemas from the structure definition and these schemas can (and should) be generated programmatically. Indeed, there are now freely available tools that do this.

Note that SDMX-EDI has just one syntactical form for the data set – there is no equivalent to a data structure specific SDMX-EDI representation. SDMX-ML also has an equivalent "one schema fits all" for the data (and also one for the reference metadata).

As all of these syntactic forms are created from the same Information Model, it is possible and practical to transform a data set or metadata set from one form to the other. Freely available tools also exist for these transformations between equivalent XML expressions, and between SDMX-EDI and SDMX-ML.

SDMX also has the concept of a registry, which is described below. Data and Metadata Sets can be registered in the registry and can be "discovered" by applications querying the registry.

3 Areas of Commonality

3.1 Concepts, Data Elements, and Value Domains

Both standards support the definition of concepts (Data Element Concept in ISO 11179). ISO 11179 associates these Concepts to actual Data Elements but it is not the role of ISO 11179 to build these Data Elements into structures, such as Data Sets and Questionnaires. This is left to other models which are domain specific, such as the CMR and SDMX.

Whilst the CMR is directly linked to the ISO 11179 model to give a conceptual structure to support the semantics of some CMR artefacts, the SDMX model is not linked in this direct way. It will be seen later that the SDMX model is actually a metamodel for defining data and metadata structures, and it is the instances of this metamodel (i.e. specific data or metadata structure definitions) that need to be mapped to ISO 11179. Therefore, in order to achieve this mapping there must be a set of rules based on the metamodel but applied to the instance. This will be explained more fully below.

3.2 Registry

Both standards support the concept of a registry

- ISO 11179 specifies a registry metamodel
- SDMX does not specify a registry metamodel
- ISO 11179 does not specify registry interfaces based on the registry model
- SDMX specifies registry interfaces based on the SDMX model

Implementation of a registry in SDMX and mapping the SDMX Information Model to a registry model is left to the implementor. Therefore, a registry developer can use an ISO 11179 registry implementation, ebXML registry implementation, or bespoke registry implementation.

We will now look at these areas of commonality in more detail.

4 Overview of the ISO 11179 Conceptual Domain

The diagram below is high level overview of the part of ISO 11179 concerned with semantics.

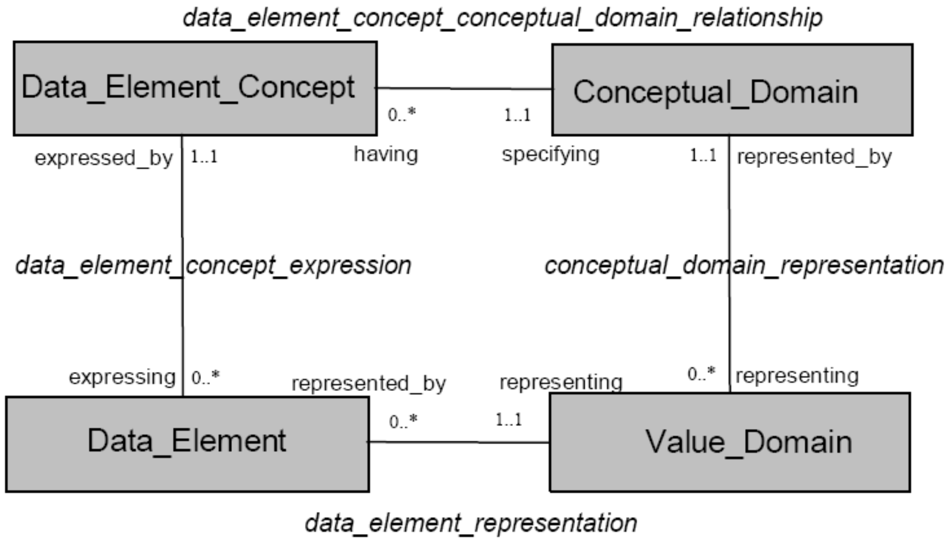


Figure 3: High level overview of the structures to specify data semantics ISO 11179

The Data_Element_Concept is a combination of concepts and properties that describe something of interest in the real world. The Data_Element_Concept is expressed independent of any internal or external representation. The metadata objects in this region are Object Classes (encompassing Concepts and Concept Relationships) and Properties, which may be combined to form Data Element Concepts.

The realm of possible values that the Data_Element_Concept can take when it is used as a Data_Element is defined in the Conceptual_Domain.

Data_Elements are derived from Data_Element_Concepts and are linked to exactly on Value_Domain.

In order to give an ISO 11179 semantic to SDMX objects, it is important to name those parts of the SDMX model (the Object Classes in the SDMX model) in ISO 11179 terms.

5 Describing CMR Object Classes

The CMR is a conceptual model that supports the survey lifecycle. “For the CMR model, the survey life cycle is comprised of the following stages (LaPlant, et al, 1996): Content, Planning, Design, Collection, Processing, Analysis, and Dissemination. Metadata to support and describe all the stages is accounted for in the CMR model” (Daniel Gillman “CORPORATE METADATA REPOSITORY (CMR) MODEL” EU Metanet R&D Project Stockholm 2002).

Element and Concept semantics are provided by an association to ISO 11179 artefacts as exampled in the diagram below.

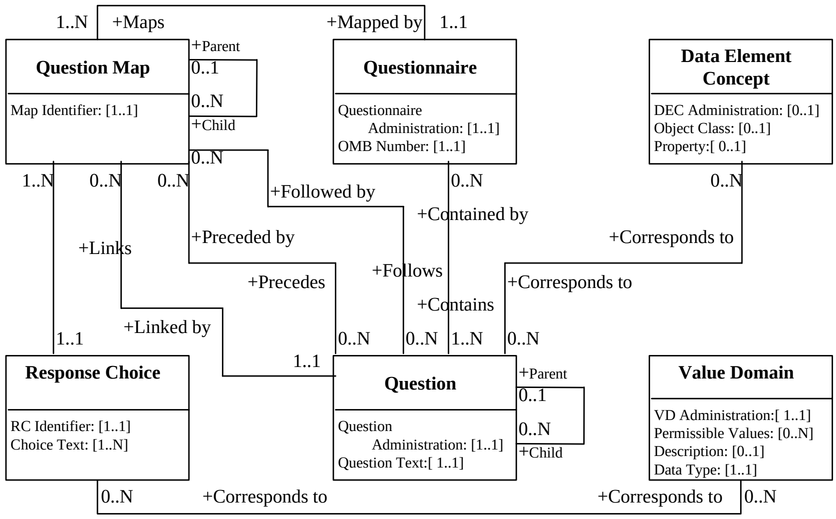


Figure 4: Example of integration of some parts of the CMR with ISO 11179 constructs

Here one can see:

- the Question is associated to Data_Element_Concept
- the Response_Domain is associated to Value_Domain

The Data Set in the CMR contains the output from a survey. Data Set components (called Data Details in CMR) are associated directly to Data_Elements in the ISO 11179 model. As the CMR Data Set is concerned with Data_Elements the classes in the ISO 11179 can be used directly in this way to give the semantics required.

6 Describing SDMX Object Classes

6.1 Introduction

The core of the SDMX Information Model is concerned with the structure of data and metadata sets. The Information Model has a far wider scope than this (mapping, process, registry), but for the purpose of this section of the paper the focus is on this core of data and metadata sets.

For the data side of SDMX, there is a very tight constraint on the Object Classes for which metadata or data can be reported, whereas for the metadata side of SDMX, metadata can be reported for any Object Class in the model, provided it can be identified (i.e. it must be an IdentifiableArtefact in SDMX terminology). It is useful, therefore, to start with the metadata set as ISO 11179 equivalent Data_Elements can be created in a semantically meaningful way.

6.2 Metadata Set Object Classes

The structure of a metadata set in SDMX is described in terms of:

- The Object Class to which metadata is to be attached, and the components that comprise the unique identifier of the Object Class
- The Concepts and the relevant Value Domains that are used to describe the metadata that can be attached to an Object of that Object Class.

A schematic of this part of the SDMX metamodel is shown below.

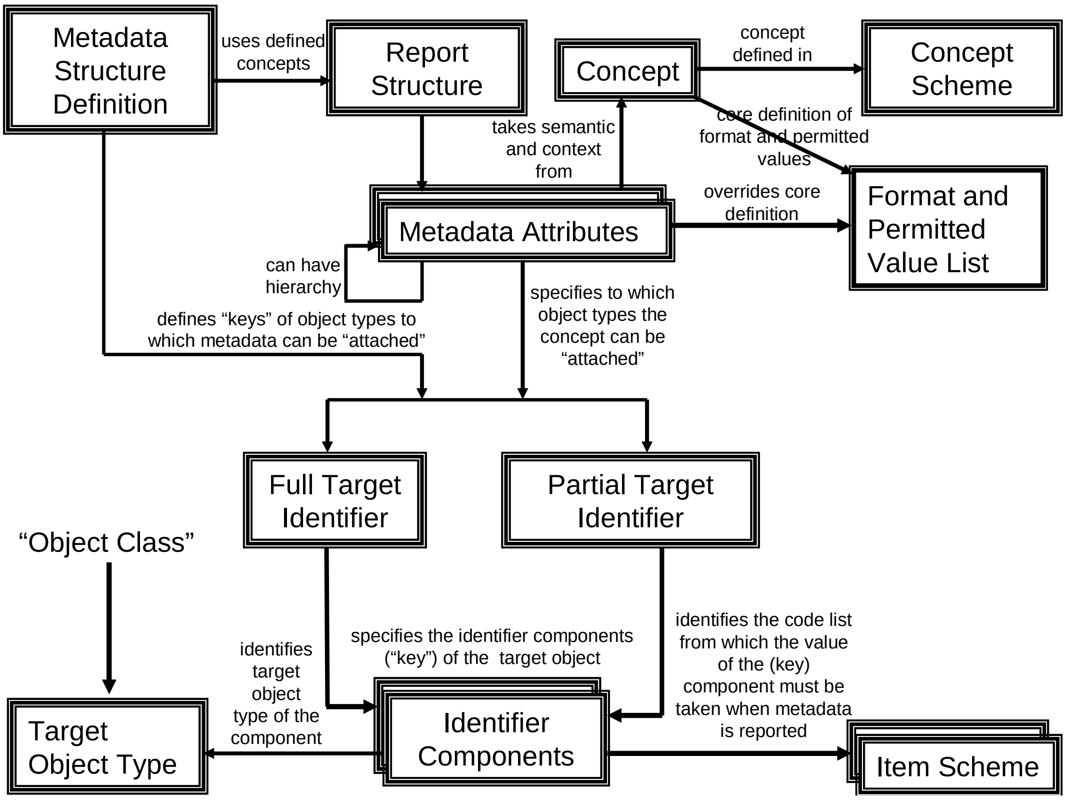


Figure 5: Schematic of the SDMX metadata structure definition constructs

In theory the metadata structure definition there is no restriction on the type of Object Classes for which metadata can be reported – in reality metadata could be reported for objects corresponding to any Object Class in any model, though in SDMX this is normally confined to the Object Classes in the SDMX model. The Object Class is identified by the sum of its Target Object Types.

Therefore in ISO 11179 terms an instance of the structure in terms of the Object Class and the Metadata Attributes can be defined as a number of Data_Elements. For instance, an organisation may wish to collect metadata from countries about the way data are collected for different domain categories such as Balance of Payments, National Accounts, Labor (e.g. the IMF SDDS). A structure definition for this type of metadata would have, as its target object, the union of a Domain Category and a Data Provider (Country). This identifies the object.

A schematic of this part of the model is shown below.

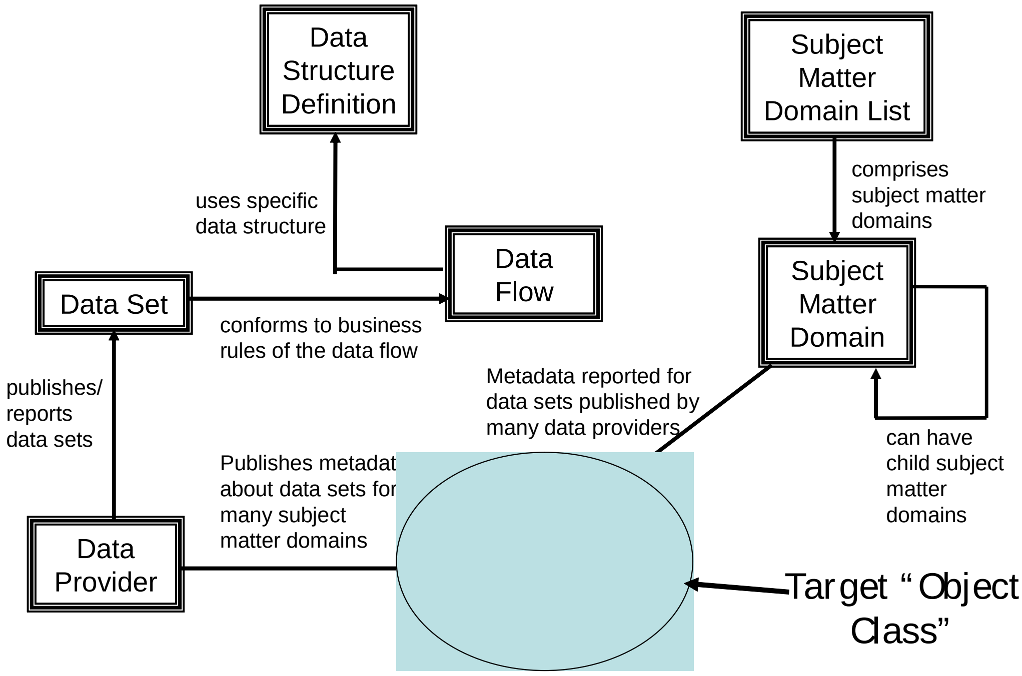


Figure 6: Schematic of part of the SDMX metamodel for data and metadata reporting

The actual metadata is collected for a variety of Concepts, each of which has a specific type of representation. Example of such Concepts are:

- Confidentiality
- Survey Source
- Simultaneous Release

The metadata structure definition would contain the following specification

Target Object Type	**Metadata Attributes**
DATA_PROVIDER_DOMAIN Identified by components: DATA_PROVIDER SUBJECT_MATTER_DOMAIN	CONFIDENTIALITY SUREVEY_SOURCE SIMULATANEOUS_RELEASE Each of these has a specific representation(Value Domain) defined in the metadata structure definition

It is possible to generate ISO 11179 Data_Elements from the instance of the metadata structure definition. Example element names would be:

```
Data_Provider_Domain.Confidentiality.Code
```

```
Data_Provider_Domain.Survey_Source.Text

Data_Provider_Domain Simultaneous_Release.Text
```

If one analyses the structure for these Data Element names in terms of the ISO 11179 model each one is structured as follows:

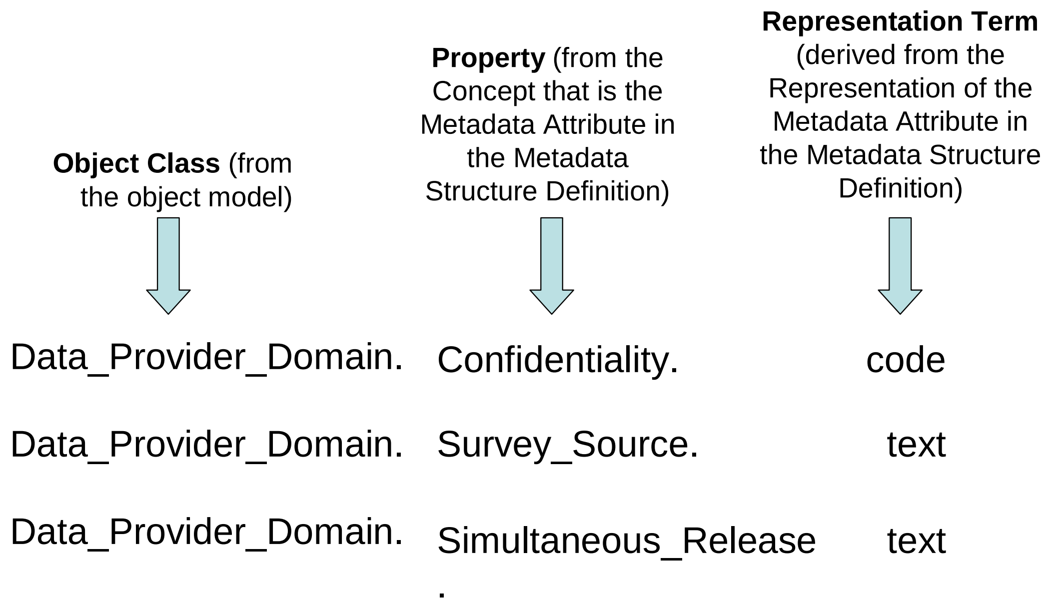


Figure 7: Derivation of data elements from an SDMX metadata structure definition

6.3 Data Set Object Classes

For the data side of SDMX the schematic of the model is:

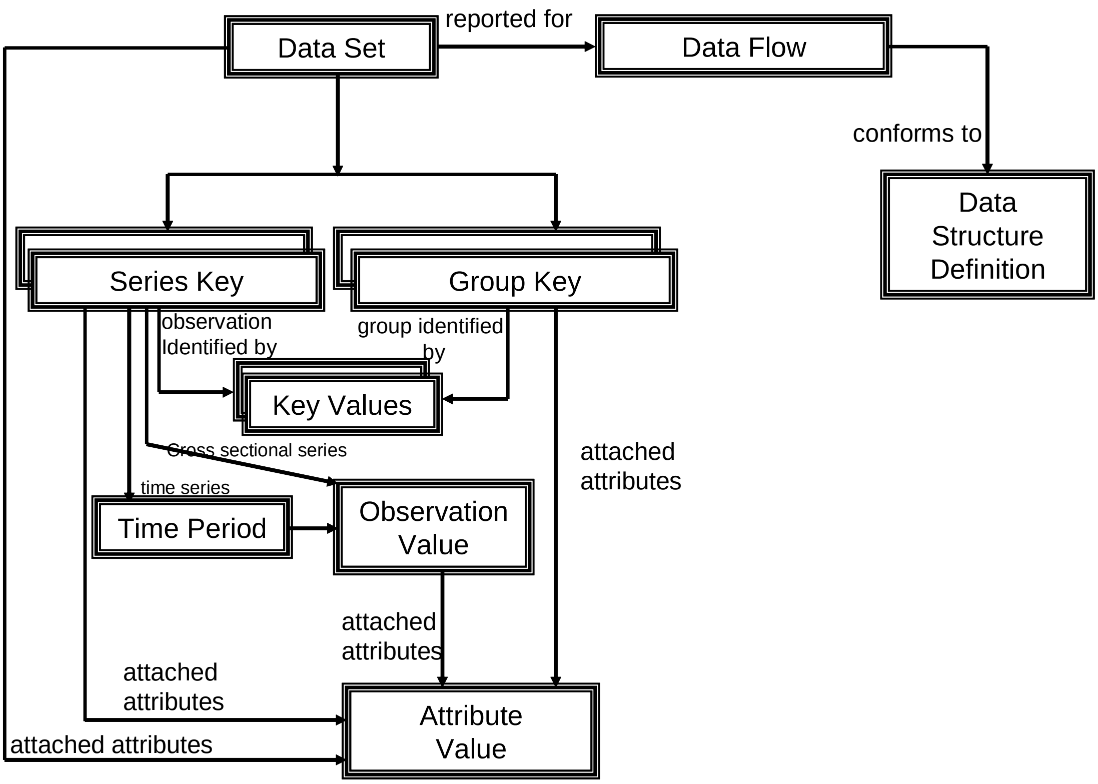


Figure 8: Schematic of the SDMX data set constructs

The object classes to which metadata (Attributes) can be attached are restricted in the model to:

- Series
- Group (qualified by the Id of the Group)
- Observation
- Data Set

The only Object Class to which a data value can be attached is the Series, qualified by the Time Period for a timeseries. The Object Class for this is the Observation Value. Therefore, the Observation Value can have both a data value and attributes.

Example ISO 11179 equivalent Data_Elements that could be generated from an instance of a Data Structure Definition would be:

```
Observation.Confidentiality.code

Observation.Value.number

Series.Availability.code

Group.Title.text

Data_Set.Title.text
```

If one analyses the structure for these Data Element names in terms of the ISO 11179 model each one is structured as follows:

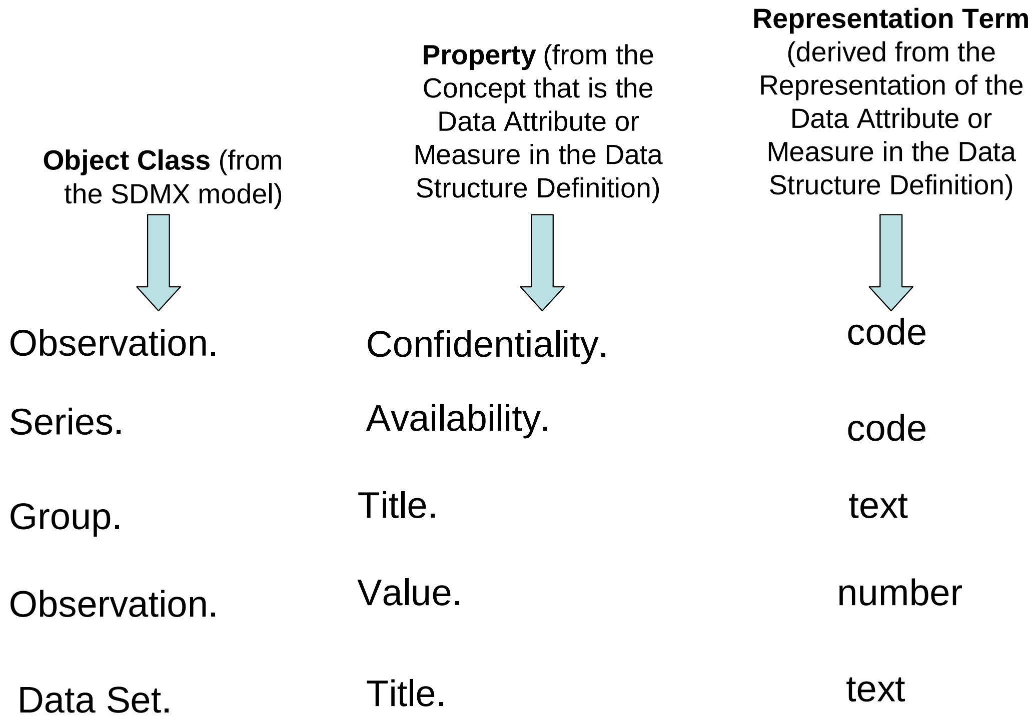
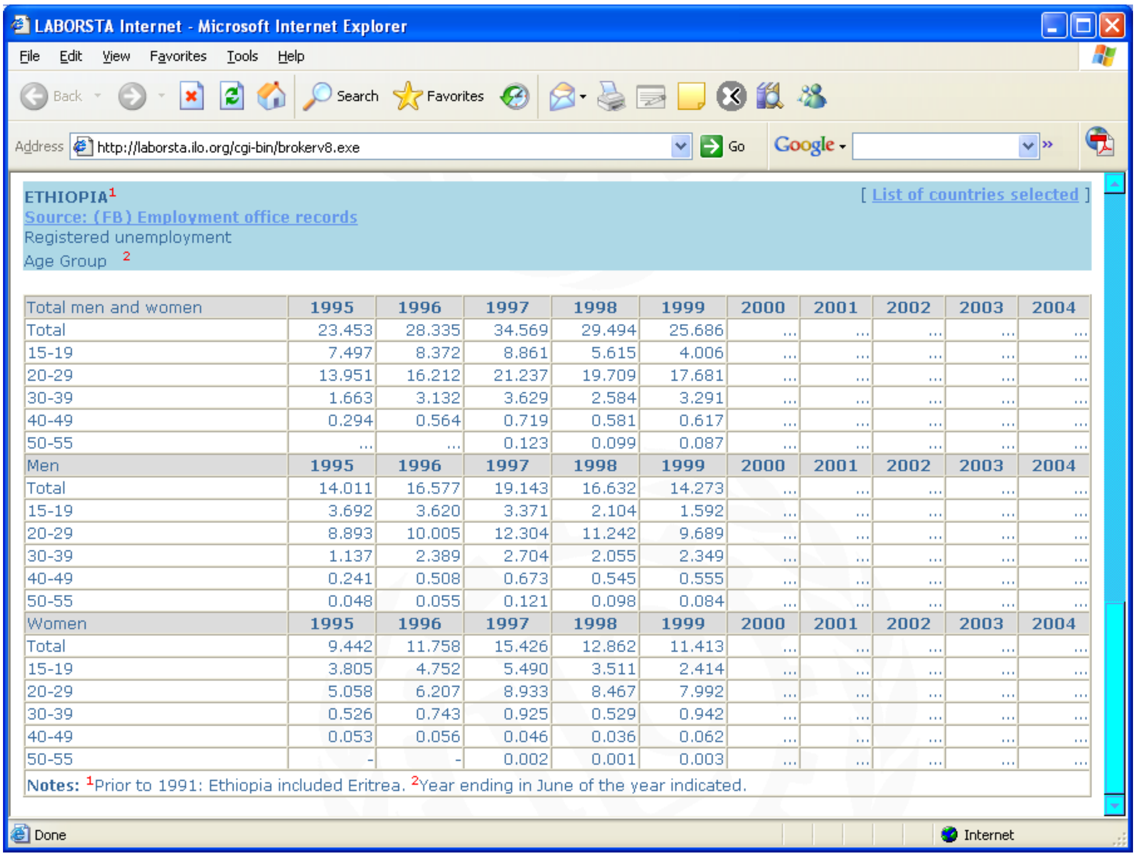


Figure 9: Derivation of data elements from an SDMX data set

This works well for the Data Attributes for the observation and data set (e.g. Confidentiality, Title, Availability) but it does not reveal the true semantic of the Observation Value, or Data Attributes for the Series or Group (a subset of the full key for the series).. The true semantic of this value is different for each Series and Group. Consider the following data set.



ETHIOPIA[1]

[List of countries selected]

Source: (FB) Employment office records

Registered unemployment

Age Group [2]

Total men and women	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
Total	23.453	28.335	34.569	29.494	25.686	...	...	...	...	...
15-19	7.497	8.372	8.861	5.615	4.006	...	...	...	...	...
20-29	13.951	16.212	21.237	19.709	17.681	...	...	...	...	...
30-39	1.663	3.132	3.629	2.584	3.291	...	...	...	...	...
40-49	0.294	0.564	0.719	0.581	0.617	...	...	...	...	...
50-55	...	...	0.123	0.099	0.087	...	...	...	...	...
Men	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
Total	14.011	16.577	19.143	16.632	14.273	...	...	...	...	...
15-19	3.692	3.620	3.371	2.104	1.592	...	...	...	...	...
20-29	8.893	10.005	12.304	11.242	9.689	...	...	...	...	...
30-39	1.137	2.389	2.704	2.055	2.349	...	...	...	...	...
40-49	0.241	0.508	0.673	0.545	0.555	...	...	...	...	...
50-55	0.048	0.055	0.121	0.098	0.084	...	...	...	...	...
Women	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
Total	9.442	11.758	15.426	12.862	11.413	...	...	...	...	...
15-19	3.805	4.752	5.490	3.511	2.414	...	...	...	...	...
20-29	5.058	6.207	8.933	8.467	7.992	...	...	...	...	...
30-39	0.526	0.743	0.925	0.529	0.942	...	...	...	...	...
40-49	0.053	0.056	0.046	0.036	0.062	...	...	...	...	...
50-55	-	-	0.002	0.001	0.003	...	...	...	...	...

Notes: [1]Prior to 1991: Ethiopia included Eritrea. [2]Year ending in June of the year indicated.

Figure 10: Example data set

This has the following (simplified) structure.

Target Object Type	**Properties (Data Attribute) or (Measure)**
Series Identified by components FREQUENCY COUNTRY GENDER AGE_GROUP	[No series level attributes in this example] TIME.NUMBER_OF_UNEMPLOYED
Country Group Identified by components COUNTRY	GEOGRAPHICAL_COVERAGE TEMPORAL_COVERAGE
Observation Value Identified by component: NUMBER_OF_UNEMPLOYED	MEASURE_UNIT UNIT_MULTIPLIER STATUS CONFIDENTIALITY

In the example data set the actual semantic of the observation value is the value of each component of the Key plus the qualifier of the Time Period as shown in the examples below.

Observations

```
Annual.Ethiopia.Male.15-19.1995.Number_of_Unemployed.number
Annual.Ethiopia.Male.15-19.1996.Number_of_Unemployed.number
Annual.Ethiopia.Male.20-29.1995.Number_of_Unemployed.number
Annual.Ethiopia.Male.20-29.1996.Number_of_Unemployed.number
Annual.Ethiopia.Female.15-19.1995.Number_of_Unemployed.number
Annual.Ethiopia.Female.15-19.1996.Number_of_Unemployed.number
Annual.Ethiopia.Female.20-29.1995.Number_of_Unemployed.number
Annual.Ethiopia.Female.20-29.1996.Number_of_Unemployed.number
```

In other words, in ISO 11179 terms each observation value is a Data_Element. Whilst statistical systems do not use this notation explicitly for aggregated data, they do nevertheless store and manipulate data with this semantic, and could, if it is useful, generate such element names.

The other Data_Elements that are derived from this specific data set example are:

Country Group Attributes

```
Ethiopia.Geographical_Coverage.text
Ethiopia.Temporal_Coverage.text
```

Observation Attributes

```
Number_of_Unemployed.Measure_Unit.code
Number_of_Unemployed.Unit_Multiplier.code
Number_of_Unemployed.Status.code
Number_of_Unemployed.Confidentiality.code
```

7 Registry

7.1 ISO 11179

ISO 11179 has an explicit registry metamodel as part of its model. Indeed the Part 3 of the standard is called “Registry metamodel and basic attributes”. A part of the registry metamodel is shown below. ISO 11179 specifies which artefacts from the model are an “Administered_Item” and can therefore be registered in an ISO 11179 registry (e.g. Data_Element, Data_Element_Concept are both an Administered_Item).

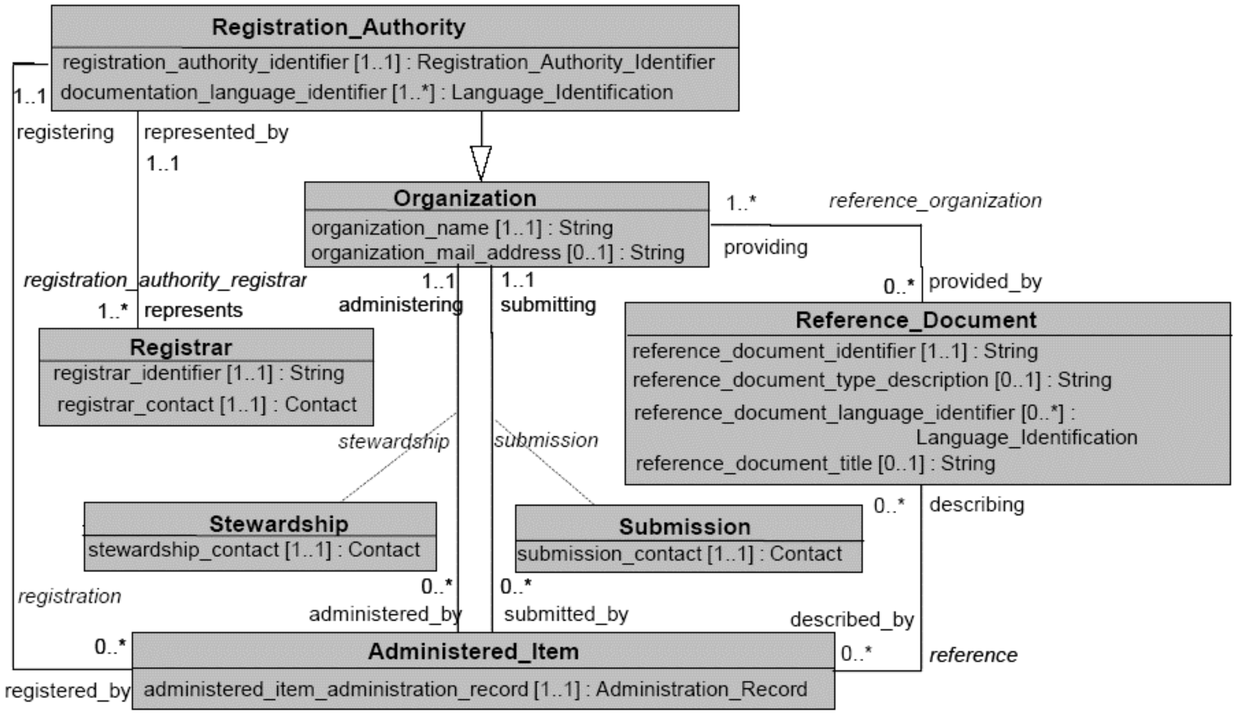


Figure 11: Class diagram of a part of the registry area of the ISO 11179 model

However, as yet there is no specified interface for interacting with an ISO 11179 registry and no registry service specification.

7.2 SDMX Registry

On the other hand, SDMX has a registry interface specification, but does not specify a registry metamodel. The SDMX registry interfaces are based on the semantics and use cases of the SDMX Information model.

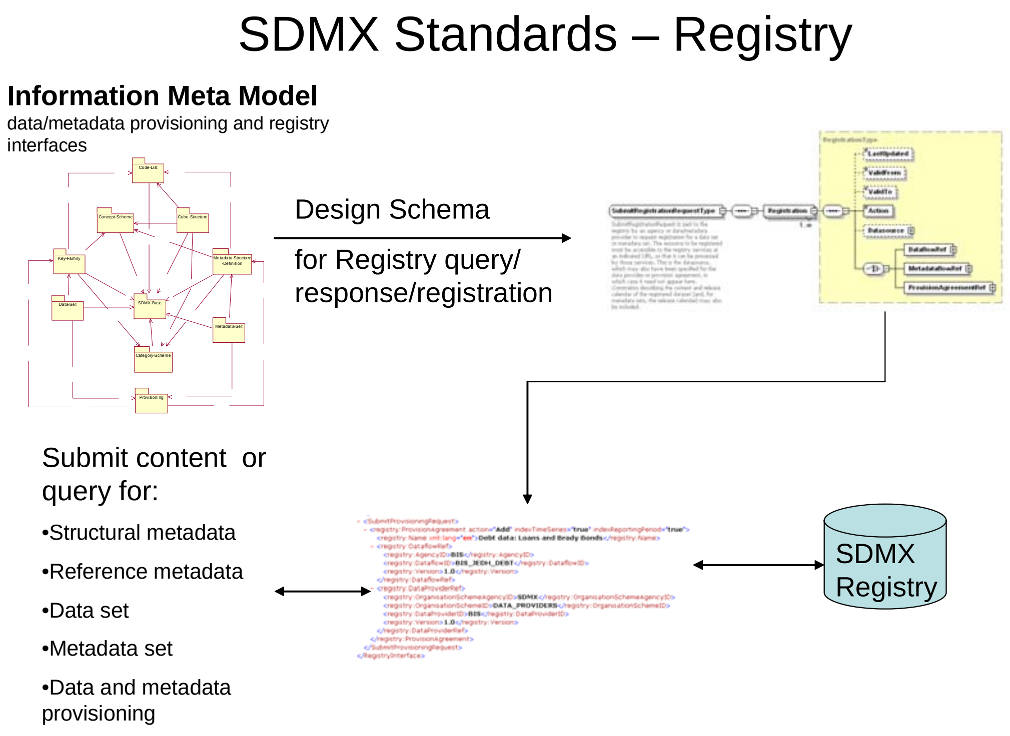


Figure 12: Deriving SDMX registry interfaces from the SDMX model

The interfaces (e.g. Submit_Registration_Request) are specified in the SDMX registry services schema. A registry provider can implement these interfaces in whatever registry product he chooses to use. The SDMX pilot project registry implementation uses a registry that complies to the ebXML registry specification (this is now an ISO standard – ISO 15000 parts 3 and 4). Of interest is that the ISO 11179 model was one of the inputs to the ebXML RIM (registry information model) and so has much functional equivalence to the "registry" region of the ISO 11179 conceptual model.

8 Summary and Conclusions

8.1 Summary

The ISO 11179 model supports the semantic specification of data. This model can be used as the basis for organising the semantics of data in an organisation. The organisation may wish to use instances of these semantic constructs to populate structural metadata about the objects that are of interest to it. For survey metadata one such structural model is the CMR and relevant constructs in this model can be associated with the semantic constructs in ISO 11179 in order provide support for the semantic meaning. For aggregated data and reference metadata there is the structural metamodel of SDMX. SDMX is a metamodel and as such the semantics of the objects are embedded in the instances of the data or metadata structure metamodel. In other words, the semantic of SDMX objects can be derived from a specific data or metadata structure.

The reference metadata metamodel supports the definition of properties (called metadata attributes in SDMX) for any identified Object Class (called Target Object Type in SDMX). This maps well to the scope of ISO 11179 and so these objects and properties can be named and stored in an ISO 11179 compliant registry where other semantic aspects can be added in accordance with the ISO 11179 model.

On the other hand, it is more difficult to realise the true semantic of the objects described in an instance (specific data structure definition) of the data metamodel. This is because the equivalent "Target Object Type" or Object Class, is a multi-faceted artefact – it is a multi-

dimensional key. Therefore, whilst the constructs of a data structure definition can realise ISO 11179 data elements, this may not be very useful to organisations. It may be that a more useful semantic can be realised by the instance of the data structure definition i.e. the data set itself. In other words, the true semantic of an observation value in an aggregated data set is realised by the sum of the actual values of the individual key components.

8.2 Conclusion

In this paper we have shown the broad scope of the three models of ISO 11179, CMR, and SDMX, in the areas in which they interact. We conclude that ISO 11179 can act as the pivotal model for mapping the semantics of CMR structures with the semantics of SDMX structures. We have done this from a conceptual point of view based on the structures of the models and the way these structures are used.

We also conclude that ISO 11179 was not built to support the definition of data elements in a metamodel because the definition of these semantics is not useful. Rather, we conclude that any definition must be derived from an instance of the metamodel. For a specific metadata structure definition this results in meaningful data elements that can be given further semantic in an ISO 11179 centric world.

However, for a specific data structure definition this may not result in meaningful data elements, and it may be necessary to base data element definitions on the content of the data set itself. We do not seek and, indeed are not qualified, to prove that any such mapping is useful for specific applications - the utility of this is for others to judge. We have established that a clear semantic mapping is possible, however, should this prove useful in an application context.