Joint UNECE/Eurostat/OECD Work Session on Statistical Metadata (METIS)
Lisbon, 11 – 13 March, 2009

Metadata Common Vocabulary: a journey from a glossary to an ontology of statistical metadata, and back

Sérgio Bacelar[1]
(Statistics Portugal)

SDMX (Statistical Data and Metadata Exchange) is not just a standard format for data exchange but it is in reality a set of technical and statistical standards and guidelines, IT architecture and IT tools with the objective of exchange and sharing of statistical data and metadata, efficiently.
SDMX Content-Oriented Guidelines (COG) recommend practices for creating interoperable data and metadata sets using SDMX technical standards, focusing on the harmonization of specific concepts that are common to a large number of statistical domains. As well as commonly understood *data structure definitions* (DSDs) allows mapping the exchanged data messages from and to internal statistical databases and systems, also commonly understood *metadata structure definitions* (MSDs) allow mapping metadata from and to statistical information systems that may be technologically or linguistic diverse. To establish these common structures, cross-domain concepts that are commonly used in SDMX messages have been identified.

One of those guidelines, Metadata Common Vocabulary (MCV) is a glossary of terms (concepts) and related definitions used in structural and reference metadata of international organisations and national data producing agencies. This glossary is a part of the Content-Oriented Guidelines which are composed by the MCV, Cross Domain Concepts (a subset of MCV) and Statistical Subject-matter Domains. The last version (2009) includes 397 terms.

The purpose of the Metadata Common Vocabulary (MCV) is to provide users with a uniform understanding of standard statistical metadata concepts. To accomplish this goal

[1] Sérgio Bacelar (sergio.bacelar@ine.pt), Metadata Unit, Methodology and Information Systems Department, Statistics Portugal.

of semantic univocity it should have been necessary to previously develop a design of a conceptual model of the respective domain, before building a glossary like MCV. As this task was not done, now it is only possible trying to reveal, ex post facto, the implicit conceptual system, detecting eventual structural inconsistencies, redundancies or incompleteness. This is not an easy task since the last version of MCV is a simple flat list of terms (non-hierarchic) and definitions in which the relations linking those terms are not transparent. More precisely, in the 2009 version of the MCV, we have for each term a set of "related terms" but we ignore what is the type of the relation between those related terms. From the previous version, the distinction between "narrower" and "broader" terms has been excluded.

In November 2008, Eurostat started an action called ESSnet on SDMX in which a group of Member States should accelerate the implementation and contribute to further development of SDMX. One of the several activities contributing to these general objectives is the further development and improvement of the SDMX Content-oriented Guidelines. Portugal coordinates the ESSnet on SDMX activity, performed by a group of NSIs from 7 countries and made a Work Package proposal in this domain (MCV Ontology).

This proposal results from previous work of Statistics Portugal with the MCV as a member of the Metadata Task Force on SDMX created by Eurostat.
We have worked with a draft version of MCV. As a first try to revise MCV we have started with the existent terms and definitions in the glossary and we have created semantic relations between terms based on the definitions of the MCV terms. This was a bottom-up or middle-out strategy. Our aim was to reveal the latent conceptual system, finding redundancies, inconsistencies, omissions, terms belonging to other domains different from statistical metadata (some of these cases were justified by the complex and interdisciplinary nature of metadata).

Concept Maps

To find omitted terms (mainly those that are more important and relevant), it is necessary to analyse the definitions of the concepts. Bearing this in mind we built a "Concept Map" representing a sample of about 20% of the terms in MCV, using *IHMC CmapTools* computer program. A concept map is a diagram showing the relationships among terms/concepts (e.g."*Classification* has *Classification changes*"). Concepts are connected with labelled arrows, in a downward-branching hierarchical structure.
Since there is a great number of terms and relations in the glossary, he visualisation of this network graphic has turned to be very difficult,

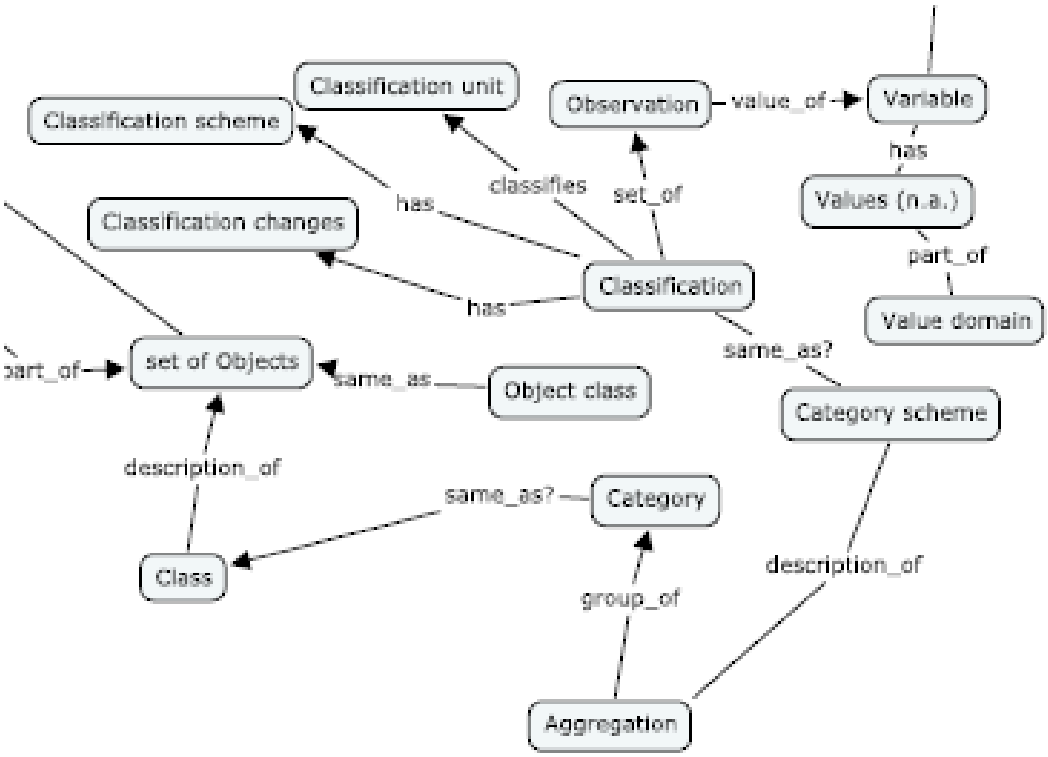

Fig.1: Concept Map

Concept_1	**relation**	**Concept_2**
Accessibility	characteristic_of	Quality
Accounting basis	type_of	Methods / procedures /conventions
Accounting conventions	same_as	Accounting basis
Accuracy	characteristic_of	Quality
Adjustment	type_of	Compilation practices
Adjustment methods	same_as	Adjustment
Administrative data	has_a	Administrative source
Administrative data	type_of	Data
Administrative data collection	collection_of	Administrative data
Administrative item	part_of	Administrative record
Administrative record	part_of	Administrative data
Age	attributeOf	Person
Agency or organization	typeOf	Analytical unit

Fig.2: Terms and relations between MCV terms/concepts (exported from IHMC CMapTools)

Using RDF

To represent semantic relations between concepts, besides Concept Mapping, the Resource Description Framework (RDF) is also to consider.
RDF is a framework for representing information in the Web, a general method to decompose knowledge into small pieces, with some rules about the semantics, or meaning, of those pieces. The point here is to have a simple method to express any fact, and yet so structured that computer applications can do useful things with knowledge expressed in RDF. RDF is particularly concerned with meaning and is designed to represent knowledge and not data in a distributed world.

The underlying structure of any expression in RDF is a collection of triples, each one consisting of a subject, a predicate and an object. A set of such triples is called an RDF graph. Each triple (e.g. "*MetadataExchange* is-a DataAndMetadataExchange") represents a statement of a relationship between the things denoted by the nodes that it links. Each triple has three parts:

1. a subject (e.g. MetadataExchange),
2. an object (e.g. DataAndMetadataExchange), and
3. a predicate (also called a property) that denotes a relationship (e.g is-a or isPartOf).

The direction of the arc is significant: it always points toward the object. The nodes of an RDF graph are its subjects and objects.

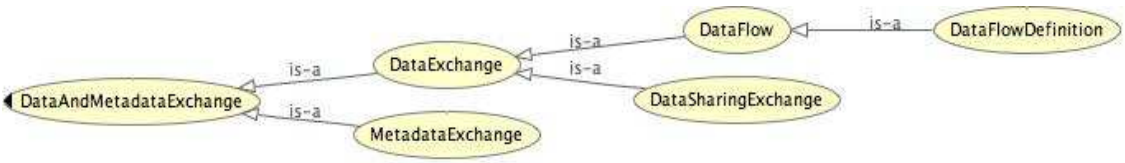

Fig. 3: RDF graph

Using SKOS

A middle range solution between RDF and an ontology, to map a glossary into a formal language, could be using SKOS - Simple Knowledge Organisation System, currently developed within the W3C framework. This specification is a bridging technology between a flat list like MCV and more rigourous logical formalism of ontology languages (like OWL). SKOS is an application of the Resource Description Framework (RDF) providing a model for expressing the basic structure and content of concept schemes such as thesauri.

Below we present a SKOS example for the MCV concept of **data**. The designation of the term, definition, label, and scope, comes from the MCV document.

```
<rdf:RDF
        .....
        <skos:Concept rdf:about=http://www.mycom/#data>
```

<skos:definition>Characteristics or information, usually numerical, that are collected through observation</skos:definition>
<skos:prefLabel>Data</skos:prefLabel>
<skos:altLabel></skos:altLabel>
<skos:related rdf:resource="http://www.my.com/#Characteristic"/>
<skos:scopeNote>Data is the physical representation of information in a manner suitable for communication, interpretation, or processing by human beings or by automatic means (Economic Commission for Europe of the United Nations (UNECE), "Terminology on Statistical Metadata", Conference of European Statisticians Statistical Standards and Studies, No. 53, Geneva, 2000).
Statistical data are data derived from either statistical or non-statistical sources, which are used in the process of producing statistical products.</skos:scopeNote>
</skos:Concept>
</rdf:RDF>

Fig. 4: SKOS code example for “data”

Ontologies

As a more complete solution to this problem, it is possible to create an ontology of statistical metadata based on the Metadata Common Vocabulary (MCV) concepts.
Once the ontology is created, the final goal is to use that ontology to create conditions to develop a new and more rigourous and structured version of the MCV.

However, an ontology is more than a glossary or a thesaurus. A thesaurus lists concepts in a hierarchical structure, while an ontology defines the relationships among these concepts more precisely, e.g. in addition to the hierarchical structure, an ontology defines similarities, differences, and other, more complex relationships.

An ontology is an explicit formal specifications of the terms in the domain (in this case in the statistical metadata domain) and relations among them. It is a model of reality in the world, usually created using an iterative design.

Since Ontology is a very strict, rigourous and formal language to represent knowledge, mapping a glossary like Metadata Common Vocabulary into a Statistical Metadata Ontology may help to reduce inconsistency, incompleteness and lack of structure in the existing MCV.

This could be done using an editing and modelling system of ontologies (e.g. Protégé which is an open source software. See in http://protege.stanford.edu). But this ontology should not be based exclusively on the terms in the MCV, since an ontology is not a mere mapping of a glossary, but as we said before, it is a representation of the world, in this case, the conceptual world of statistical metadata.
This ontology may ease inter-operability between systems accessing statistical metadata in different NSIs, by using a formal language for description and labelling statistical reference metadata, enriching by this manner the available information with machine-processable semantics.

Methodology to create an ontology

The methodology we propose to create an ontology of statistical reference metadata is based mainly on Noy e McGuiness (2001) but also on Uschold and King (1995). The main steps of the process are:

First, we have to determine the domain and scope of the ontology. That involves also to define the goals with which the ontology will be used, what type of questions will be answered using the knowledge contained in the ontology. Finally it is also important to know who will use and maintain the ontology.

Second it is also necessary consider reusing existing ontologies, importing them using some ontology developing environment (e.g. Protégé).

Third, we must enumerate the important terms in the ontology (in this case, terms coming from MCV glossary). With this purpose we have to clarify with domain experts (statistical metadata), which are the important terms and which are their properties. To accomplish this goal it is important to revise the definition of each concept, with precision and no ambiguity and to identify the terms referring to each concept and the semantic relations linking them.

Fourth, it is absolutely necessary to define the classes and the class hierarchy. There are several possibilities: a top-down approach: starting with the definitions of the most general classes and ending progressively with more specific classes. A top-down approach to the MCV, starting from existent statistical metadata ontologies (e.g. Froeschl et al. (2003) or Sundgren (2008)), or classifications of statistical metadata, would result into different perspectives of this glossary; a bottom-up approach, starting with the current terms and definitions in the glossary and looking for relations and meanings to discover the conceptual system, that is, the more general classes; and finally a middle-out approach: this is a mixture of these two methods. Beginning with a core concept, we identify the more general and specific ones. We have created four main classes for the MCV, according to SDMX Content-Oriented Guidelines: Framework classification (Draft March 2006, p.6): 1. General metadata (derived from ISO, UNECE and UN documents); 2. Metadata describing Statistical methodologies; 3. Metadata describing Quality assessment; 4. Terms referring to Data and metadata exchange (SDMX information model and data structure definitions, etc.). In spite of this example, we think that this classification of MCV terms is not adequate to build classes adequate to obtain a real ontology of statistical metadata.

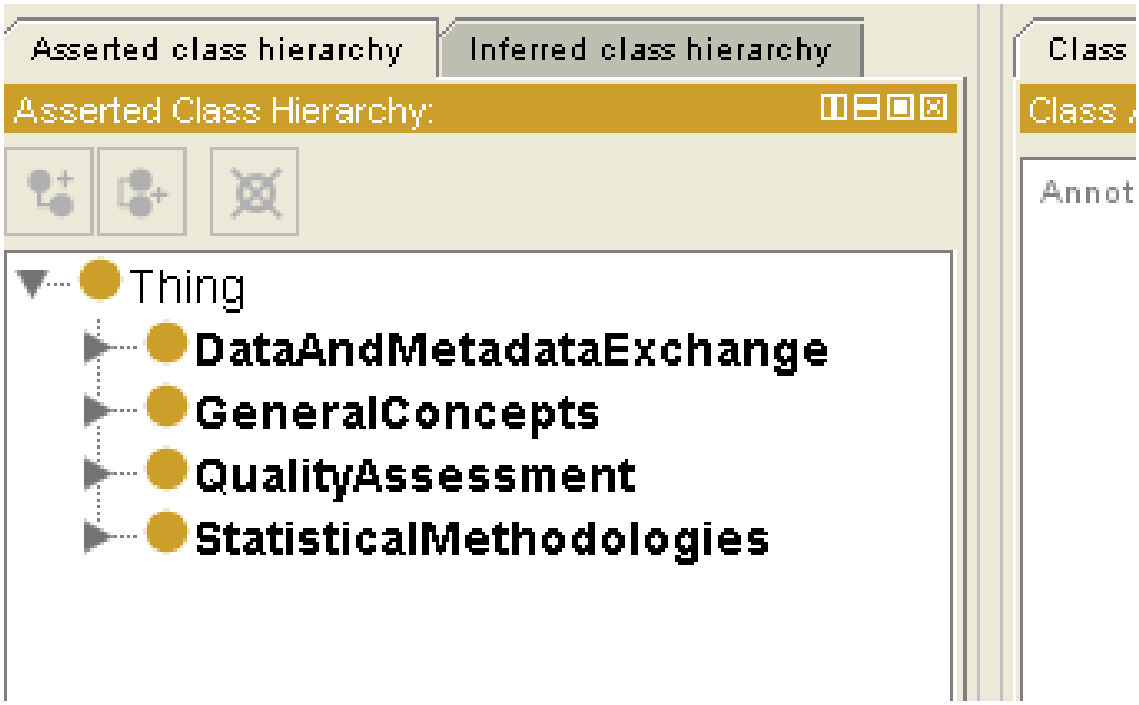

Fig. 5: Classes and subclasses (using Protégé)

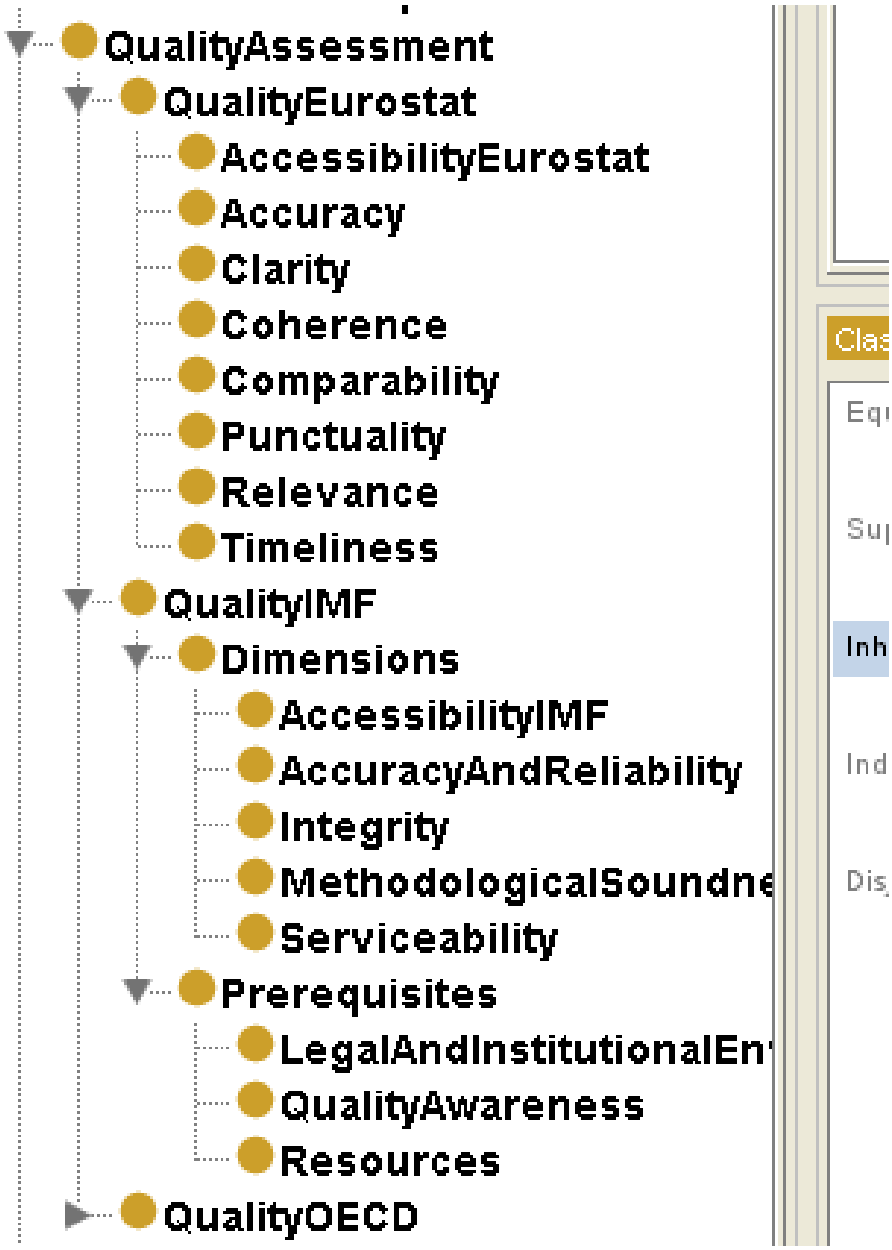

Fig. 6: Classes and subclasses: QualityAssessment (using Protégé)

Fifth, we have to define the properties of classes (slots). This fact implies not only to describe the internal structure of the concepts, but also to determine which is the class each property (slot) describes. These properties will be attached to the classes.

Sixth, we must define the facets of the slots, that is to define the additional properties related to, or necessary for properties (e.g. type values, cardinality, bidirectionality/inverse).

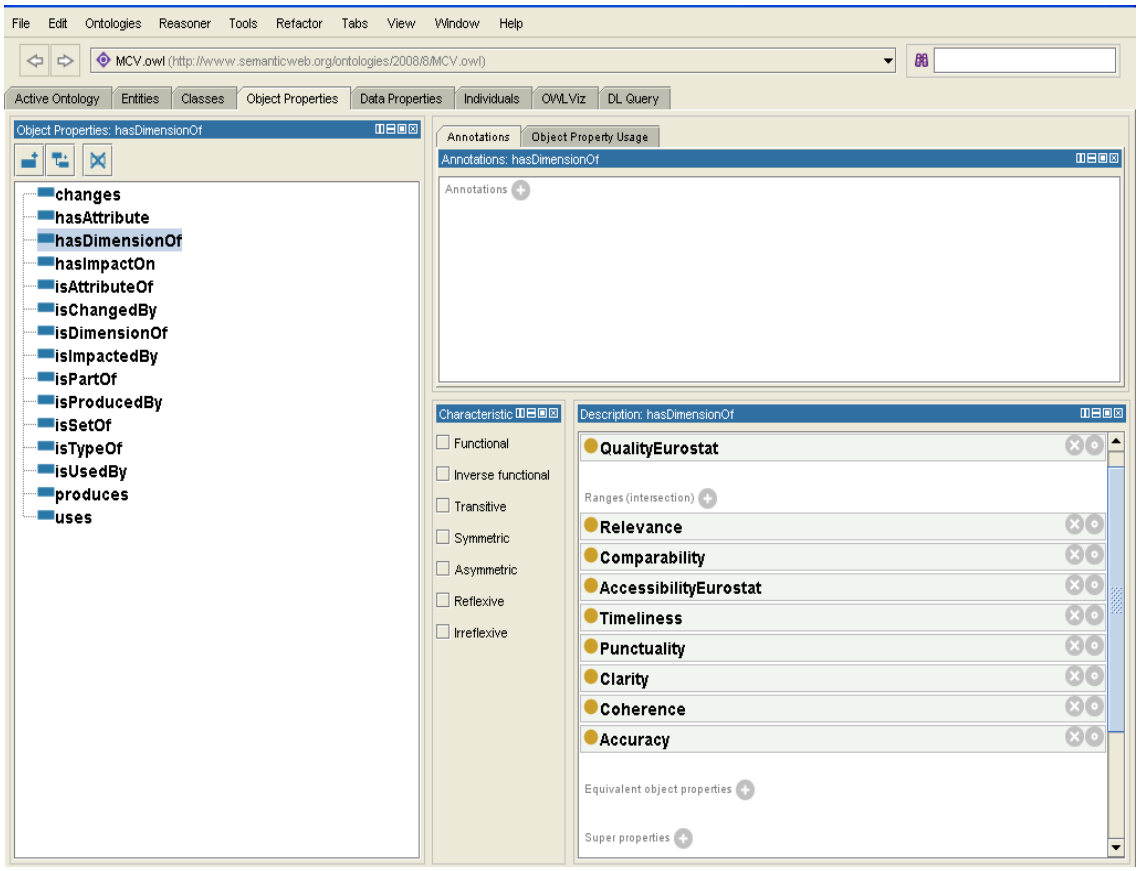

Fig.7: the class **QualityEurostat** (quality according to Eurostat) hasDimensionOf (**property**) called **relevance**

Seventh, it is necessary to create instances of the classes. With this purpose we have to choose a class and then create an individual instance of that class and fill the slot values (e.g. *Portuguese Labour Force Survey* of the second quarter of 2004).

Eighth, the knowledge of statistical metadata should be codified, that is the representation of the conceptual system should be done using a formal language (e.g. Web Ontology Language – OWL). Of course this code can be generated by the ontology editor, using an OWL plug-in.
The following shows an example of this OWL code generated by Protégé:

```
……..
<owl:Ontology rdf:about="">
      <rdfs:comment
         >Metadata Common Vocabulary (MCV) ontology.</rdfs:comment>
   </owl:Ontology>
……..
 // Object Properties
<!-- http://www.semanticweb.org/ontologies/2008/8/MCV.owl#uses -->

   <owl:ObjectProperty rdf:about="#uses">
      <owl:inverseOf rdf:resource="#isUsedBy"/>
   </owl:ObjectProperty>
……..
// Classes
```

```
 <!--
http://www.semanticweb.org/ontologies/2008/8/MCV.owl#ComputerAssistedInterviewing --
>

    <owl:Class rdf:about="#ComputerAssistedInterviewing">
       <rdfs:subClassOf rdf:resource="#DataCollection"/>
    </owl:Class>
```

Fig 8: Codification: Ontology Web Language.

We can see in this code essentially two statements: first, that “uses” is an object property that has an inverse property called “is used by”; second, that “Computer Assisted Interviewing” is a subclass of “Data Collection”.

Eighth, the evaluation of the ontology will be based on specific requirements and competency questions. These questions model common ‘real world’ queries that actual users of the ontology would state with high frequency. To allow the system to perform an ontology reasoning it is essential to provide tools and services (reasoners) to help users answer queries over ontologies and classes and instances, e.g., to find more general/specific classes or to retrieve an individual matching an existing query. The following proposition is an example of an ontological query: "Is there any survey with quarterly frequency that uses any classification system and has a dissemination format as an on-line database?"

Finally this process has to be documented, writing documentation containing every decision made by the people that have created the ontology.

Conclusion

With the aim to improve the quality of the MCV glossary, we tried to specify it using different formal languages. We started using a graphical visualisation scheme like Concept Maps. That was a first attempt to represent semantic (meaning) relations between the terms/concepts in the MCV. This solution rapidly proved to be very difficult to read or use, since in this glossary there is a huge amount of terms. As a simple graphical language, Concept Maps could not be read by computers making impossible any inter-operability between systems.
Using RDF or one of the RDF applications like SKOS, tailored to represent glossaries or thesauri, could be an answer to the previous problem. This knowledge representation language could be understood by machines realising the dream of a semantic Web (a meaningful web) for searching statistical metadata associated with statistics data in the NSIs websites.
But a solution based in RDF is not able to allow for reasoning, causing difficulties in providing answers to the user queries about statistical metadata.
It seems that the best, although difficult way, to deal with these kind of problems is to create an ontology, coding the world of statistical metadata into a web ontology language. This way it should be possible to formalise the terms in a glossary so that simultaneously, it would assure rigour and structure, and it would also be machine readable.

▪

References

GRUBER, Thomas R. Toward principles for the Design of Ontologies Used for Knowledge Sharing. In: FORMAL ONTOLOGY IN CONCEPTUAL ANALYSIS AND KNOWLEDGE REPRESENTATION. March 1993. Padova. Italy. Available as Technical Report KSL 93-04, Stanford University.

NOY, Natalya F.; McGUINNESS, Deborah L. Ontology Development 101: A Guide to Creating Your First Ontology. Stanford University: Stanford, 2001. Available in: <http://www.ksl.stanford.edu/people/dlm/papers/ontology101/ontology101-noy-mcguinness.html> Accessed in: May 2006.

USCHOLD, Mike; KING, Martin. Towards a Methodology for Building Ontologies. In: “WORKSHOP ON BASIC ONTOLOGICAL ISSUES IN KNOWLEDGE SHARING”. University of Edinburgh: Edinburgh, 1995.

Software

IHMC CMapTols - http://cmap.ihmc.us/conceptmap.html
Protégé - http://protege.stanford.edu/
SKOS - http://www.w3.org/2004/02/skos/