MCV Ontology

Deliverable 1

ESSnet on SDMX - Work Package 2

WP Responsible: Sérgio Bacelar
Date: 08/2010

Document Properties

Project	ESSnet on SDMX – WP2		
Type of Document	First deliverable		
Revision:	1	**Status:**	Draft
Created by:	Goretti Nunes Isabel Morgado Luísa Saraiva Olga Mendes Sérgio Bacelar	**Submission Date:** 12.08.2010	
Reviewed by:	Christophe Roche, Rute Costa		
Approved by:			

Document Change Record

Revision	Date	Change
1		
2		
3		
4		

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

This report describes the methodology adopted to create an ontology of statistical metadata based on the Metadata Common Vocabulary (MCV) concepts, provides an overview of the first results achieved, having been produced by the working group of the *ESSnet on SDMX – Work Package 2: MCV Ontology*, with the following participants: Isabel Morgado, Sérgio Bacelar, Goretti Nunes, Luísa Saraiva, Olga Mendes, from Statistics Portugal, and by the consultants Prof. Christophe Roche (PhD) from Université de Savoie (France) and Prof. Rute Costa (PhD) from Universidade Nova de Lisboa (Portugal).

To begin with, we refer the purpose, domain and scope of the ontology we are building. Secondly we unfold our methodology of work: first steps and questions, first approach to classify the concepts, path followed to identify key concepts of the ontology and at last, how we established a set of rules to define concepts and the conventions used to represent them in concept maps. Thirdly we show the results of our work, defining a set of concepts in formal and natural language, using a top-down and bottom-up approach. Finally we evaluate the work done and forecast the work we intend to do in the next phase of the project.

PURPOSE

The final aim on building the ontology is to create conditions to develop a new version of the MCV.

> *An ontology defines a set of representational primitives with which to model a domain of knowledge or discourse. The representational primitives are typically classes (or sets), attributes (or properties), and relationships (or relations among class members).* (Gruber, 2009)

An ontology is an explicit formal specification of the concepts and relationships among them in a domain (Gruber,1993).

Since MCV has lack of structure and inconsistencies on concepts definitions, the ontology will help to reduce these problems due to its essential condition as a formal language.

DOMAIN AND SCOPE

The building up of an ontology presumes the definition of the domain of knowledge and the scope of its application as well as the identification of its users.

The broader framework of our work and goal is the statistical metadata domain. We have chosen the sub domain *data and metadata exchange* to set to work, since it is the core of the SDMX. We consider that the users of the ontology will be as varied as *data providers*, *data consumers* and/or *maintenance agencies*, thus implying that the ontology is meant to be widely known and used.

As we are dealing with concepts created to help organizations to exchange data and metadata, we consider that besides all other statistical processes that occur in a statistics

organization, our main focus is the usage of metadata within the process of *data and metadata exchange* in the context of the SDMX initiative.

METHODOLOGY OF WORK

FIRST STEPS AND FIRST QUESTIONS

Our first step was to analyse the MCV version of 2009, namely the terms and definitions, aiming to grab the underlying conceptual structure of its organization.

Initially we tried to identify concepts in the MCV based on semantic relations between terms.

This has not been an easy task, since the last version of MCV is a simple flat list of terms (non-hierarchic relations between terms) and some of the definitions of the concepts don't facilitate the task of determining the type of relations linking those concepts. There are terms that have related terms associated, but the type of relation between those related terms is not clearly perceivable.

It was becoming clear the fact that the MCV alone was not enough to develop our work, due to the complex and interdisciplinary nature of metadata.

Having a heterogeneous point of departure, resulting from different conceptual frameworks and containing concepts from disparate sources, there are three main and framing questions to move forward:

- which are the core concepts to build an ontology of statistical metadata?
- what is the specific purpose of the ontology and to whom it will be useful to?
- which are the concepts belonging to the domain we chose that should not really be part of the MCV?

CLASSIFICATION OF CONCEPTS: CRITERIA AND RESULTS

Clustering concepts is a way to implement the creation of an ontology, since this minimizes the average semantic distance between concepts inside each cluster.

We used two criteria to classify the 397 terms of the Metadata Common Vocabulary (MCV): the first criteria was inspired by the "General Statistical Business Process Model" (GSBPM), the second one resulted from the suggestion of the Content-Oriented Guidelines document about MCV's organisation.

The GSBPM provides a model that can be seen as a flexible tool to describe and define the set of business processes needed to produce official statistics and which is also a basis for statistical organizations to agree on standard terminology useful for the development of statistical metadata systems. We only used Level 1 that corresponds to the nine phases of the statistical business process. According to the Content-Oriented Guidelines metadata concepts covered by the MCV represented general concepts, concepts dealing with statistical methodologies, quality and data and metadata exchange. The working group decided to apply them according to its professional knowledge of the statistical activity.

The expected results of these classifications were the possibility of high-lightening MCV's organisation and structure, as well as the possibility to improve the perception of concepts relations, which was partially achieved. The main advantage was the

prospect of having a very complete depiction of the MCV's concepts composition and provision.

The evidence of the mixture of various elements lead us to the inevitability of building some statistics on them. The following table shows a cross-tabulation between categories of classification of the concepts by their source in absolute value and percentage. Our main conclusion is that 50% of the concepts origin is from the *source SDMX*.

Table 1 – MCV content according to the Source

	SDMX	%	ISO/IEC	%	Other	%	Total
Source	182	50	45	12	136	37	**363**
Cross-reference terms (a)	0		0		0		**34**
Total	**182**		**45**		**136**		**397**

(a) Without source.

IDENTIFICATION OF KEY CONCEPTS

MCV is a subset of Content-oriented Guidelines meant to support the implementation of the SDMX standard on the exchange of statistical information. The SDMX Information Model (SDMX-IM) is one of documents of the SDMX standard which identifies the main concepts of SDMX and their relationships, using a formal modelling language (UML) as a basis for the development of SDMX applications.

In order to better understand the class diagrams of the information model, we analysed files stored in the SDMX registry; then we focused on two main structures of the information model: *data structure definition* and *metadata structure definition*. We identified afterwards the concepts that compose each one, based on the available definitions.

This method was iterated for each of the concepts that are part of the initial concept.

DEFINITION AND REPRESENTATION OF CONCEPTS

We are considering four types of definitions that can occur on their own or gathered in a same concept:

- an *essential* definition: we define the concept by its broader genus and indicating its specific differentia (e.g. *a "Metadata Structure definition" is a "Structure" of metadata*);
- a *functional* definition: describes the function of a concept (e.g. *a "Metadata Structure Definition" has the function of producing a report about reference metadata*);
- a *descriptive* definition: lists the attributes of a concept (which qualifies the objects comprised by the concept);

- a *constitutive* definition: identifies the concepts that are part of the definition of the concept specifying the relationships that they maintain with the concept: relationships of the type "part of", "function", etc. (e.g. *a "Metadata Structure Definition" is composed of an object type, a report structure, etc.).*

The definition of each concept shall be then represented in:

- natural language definition
- formal language definition

Both, the natural language definition as well as the formal definition, represent the same concept which means that they should have the same composition.

The terms in MCV are defined in natural language, whereas concepts in ontology which represent term's meaning are defined in formal language (in general, the concept name and the term are the same).

We are developing, simultaneously, a terminological database for the natural language definitions.

The formal language definition is represented in a concept map which is a diagram showing the relationships among concepts that can be seen as a first step in the ontology-building, once it provides a "human-centred interface to display the structure, content, and scope of an ontology" (Institute for Human & Machine Cognition, n.d.).

We have been using *IHMC CmapTools* open source software to build concept maps diagrams.

CmapTools: methodological and graphical conventions

CmapTools is an editor of concept maps of the semantic networks type. It is based on a model only defined by a set of nodes and binary relationships linking the nodes between themselves. There is only one type of node, without any distinction between the types of concept, class, set or instance). As the notion of attribute is missing, they will be represented by relationships.

It is by the reasons mentioned above, that we will be following the next methodological and graphical conventions:

concept – for each concept to be defined we will build a concept map. The name of the map shall be the name of the concept. The core concept (the concept that is being defined) is framed in a different color, typeset in a bigger font and in bold. The name of the concept will be the more explicit possible, and we will not be using acronyms;

relationship – relationships link different objects between themselves; the predefined relationship of the type "is a", "part of" and "function" are represented by different colors;

attribute – attributes belong to the object itself;

The attributes and the relationships of the conceptual model will be represented by edges in *CmapTools*. We will distinguish attributes and relationships by the following graphical conventions:

- attributes are represented by *CmapTools* edges where names are prefixed by "has". For example: "has identifier";
- *CmapTools* edges representing attributes are represented by dotted directed edges (arrows);
- the value type of the attribute is represented by a *CmapTools* inside a squared dotted box.

annotation – represents the terminological record including the natural language definition, the standardized terms and the terms in use.

Besides the methodological steps mentioned above, we adopt the convention of using one concept map by definition, and one concept map by perspective.

The aim afterwards is to build this representation in Web Ontology Language (OWL).

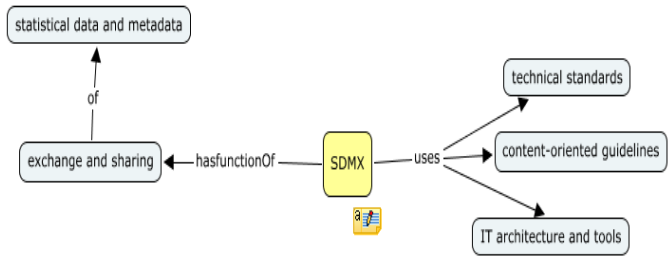
RESULTS AND EVALUATION

The methodology exposed above, was applied using first, a top-down approach, started from the definition of the general concept of "data and metadata exchange"; and the second, a bottom-up approach, started from two core concepts of SDMX: "Data structure definition" and "Metadata structure definition".

Definitions of concepts in formal and natural language

Top-down approach

Figure 1 - Statistical Data and Metadata Exchange

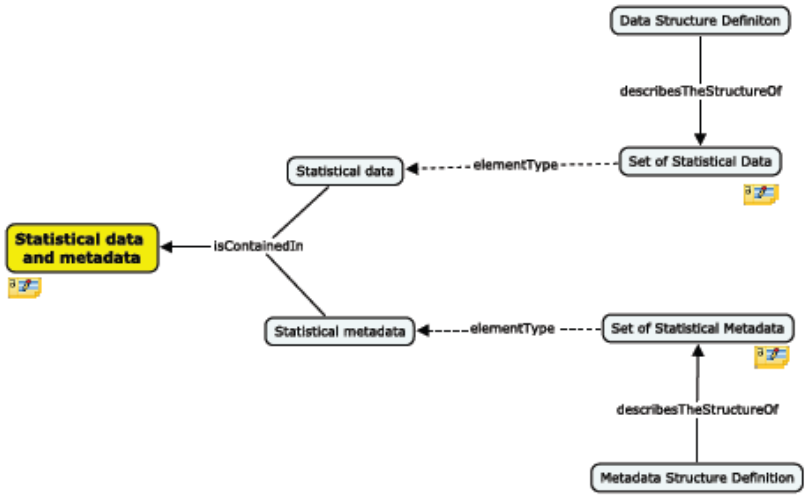


SDMX-statistical data and metadata exchange

definition: set of technical standards and content-oriented guidelines, together with an IT architecture and tools to be used for the efficient exchange and sharing of statistical data and metadata [=ISO/TS 17369:2005]

source: SDMX 2009

Figure 2 - Statistical Data and Metadata



statistical data and metadata

metadata

definition: data that define and describe other data.

source: ISO/IEC FDIS 11179-1 "Information technology - Metadata registries - Part 1: Framework", March 2004

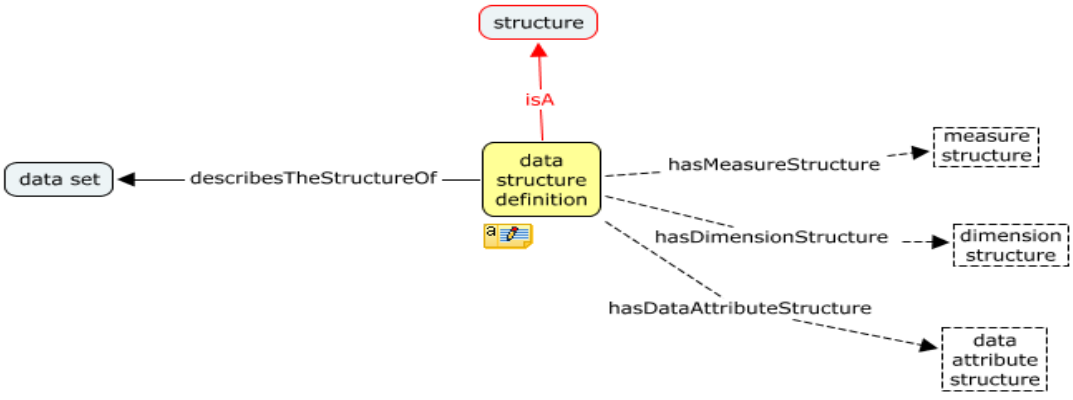
data

definition: characteristics or information, usually numerical, that are collected through observation.

source: The International Statistical Institute, "The Oxford Dictionary of Statistical Terms", edited by Yadolah Dodge, Oxford University Press, 2003

Bottom-up approach

Figure 3 - Data Structure Definition



data structure definition

definition: structure of a data set, composed by descriptor concepts organized in a measure structure, a dimension structure and a data attribute structure.

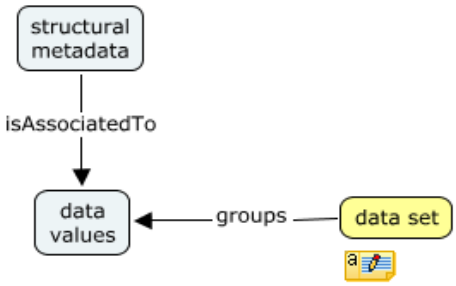
source: ESSNET, Wp2, adapted from MCV

structure

definition: pattern or model consisting of a group of elements that is specifically organized to identify and represent a set of features or categories.

source: ESSNET, Wp2

Figure 4 - Data set



data set

definition: organized collection of data values and their associated structural metadata.

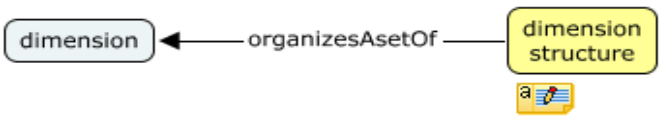
source: ESSNET, Wp2

structural metadata

definition: metadata that act as identifiers and descriptors of the data.

source: SDMX (2009)

Figure 5 - Dimension structure



dimension structure

definition: organized set of dimensions.

source: source: ESSNET, Wp2

Figure 6 - Dimension

dimension

definition: statistical concept used, in combination with other statistical concepts, to identify and describe a statistical series or single observations.

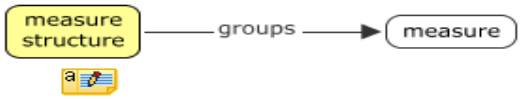
source: SDMX (2009), adapted from *User guide*

statistical concept

definition: statistical characteristic of data

source: SDMX (2009)

Figure 7 - Measure structure



measure structure

definition: organized set of measures.

source: ESSNET, Wp2

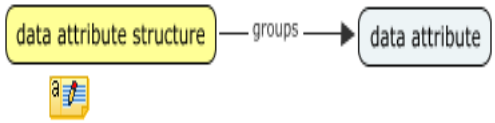
Figure 8 - Measure

measure

definition: statistical concept identifying a phenomenon for which quantitative information is provided.

source: ESSNET, Wp2, adapted from MCV

Figure 9 - Data attribute structure

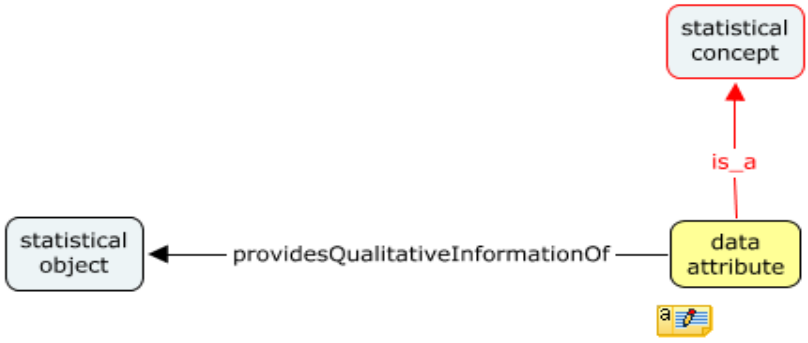


data attribute structure

definition: organized set of data attributes.

source: ESSNET on SDMX, Wp2

Figure 10 - Data attribute



data attribute

definition: statistical concept providing qualitative information about a specific statistical object.

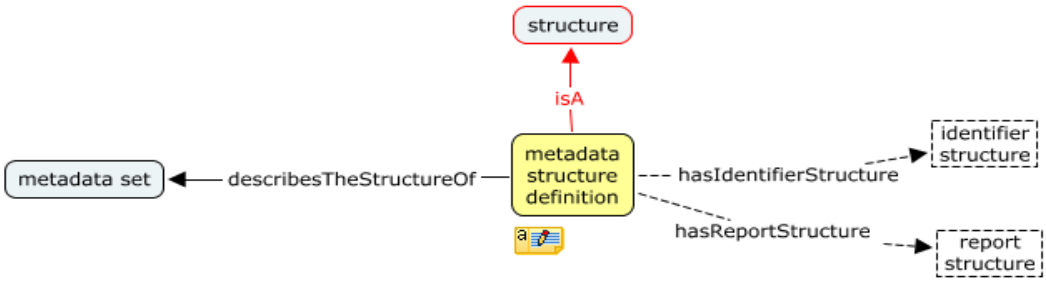
source: ESSNETon SDMX, Wp2, adapted from the context of MCV's "attribute"

Table 2 - SDMX cross-domain concepts by use (2009)

data structure definitions – data attributes

Adjustment - coded	Maintenance agency
Adjustment - detail	Observation pre-break value
Age	Observation status
Base period	Observation
Civil Status	Occupation
Comment	Originator data identifier
Compiling agency	Recording basis
Confidentiality - status	Reference area
Counterpart reference area	Reference period
Coverage	Reference period - weights
Currency	Release calendar access
Data compilation	Release policy - user access
Data dissemination agency	Reporting agency
Data provider	Sex
Data set identifier	Time format
Data update	Time period
Decimals	Time period - collection
Dissemination format - news release	Timeliness
Dissemination format - publications	Timeliness - source data
Education level	Title
Embargo time	Unit multiplier
Frequency	Unit of measure
Frequency detail	Unit of measure detail
Frequency of dissemination	Valuation
Index type	

Figure 11 – Metadata Structure Definition

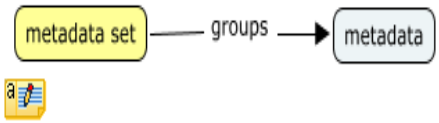


metadata structure definition

definition: structure of a metadata set, composed by an identifier structure and a report structure

source: ESSNET, Wp2

Figure 12- Metadata set



metadata set

definition: organized collection of metadata.

source: ISO/IEC Committee Draft 11179-3: 2007, Information technology - Metadata Registries (MDR) - Part 3: Registry Metamodel and basic attributes, August 2007, adapted

Figure 13 - Identifier structure

identifier structure

definition: organized sequence of metadata attributes.

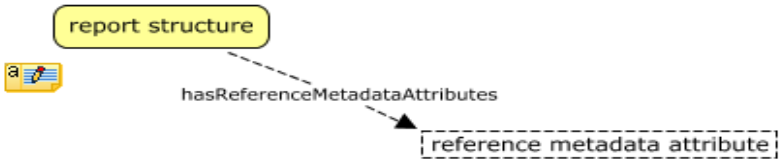
source: ESSNET, Wp2

metadata attribute

definition: statistical concept providing qualitative information about a specific statistical object.

source: ESSNETon SDMX, Wp2, adapted from the context of MCV's "attribute"

Figure 14 - Report structure



report structure

definition: organized list of reference metadata attributes.

source: Euro-SDMX Metadata Structure Definition – Message implementation guidelines

reference metadata

definition: metadata describing the contents and the quality of the statistical data.

source: SDMX (2009)

Table 3 - SDMX cross-domain concepts by use (2009)

metadata structure definitions – reference metadata attributes

Accessibility
Accuracy
Accuracy - overall
Non-sampling error
Sampling error
Adjustment
Adjustment - detail
Base period
Clarity
Classification system
Coherence
Coherence - cross domain
Coherence - internal
Comment
Comparability
Comparability - between domains
Comparability - geographical
Comparability - over time
Confidentiality
Confidentiality - data treatment
Confidentiality - policy
Confidentiality - status
Contact
Contact email address
Contact fax number
Contact mail address
Contact name
Contact organisation
Contact organisation unit
Contact person function
Contact phone number
Cost and burden
Cost and burden - efficiency management
Cost and burden - resources
Coverage
Coverage - sector
Coverage - time
Data collection
Data compilation
Data editing
Data presentation
Data description
Disseminated detail
Data provider
Dissemination format - publications
Documentation on methodology
Documentation on methodology - advance notice
Frequency
Frequency of data collection
Frequency of dissemination
Grossing / Netting
Index type
Institutional mandate
Institutional Mandate - data sharing
Institutional Mandate - legal acts and other agreements
Institutional Mandate - respondent relations
Maintenance agency
Metadata update
Metadata last certified
Metadata last posted
Metadata last update
Professionalism
Professionalism - code of conduct
Professionalism - impartiality
Professionalism - methodology
Professionalism - statistical commentary
Punctuality
Quality management
Quality assessment
Quality assurance
Quality documentation
Recording basis
Reference area
Reference period
Reference period - weights
Release policy
Release calendar
Release calendar access
Release policy - commentary
Release policy - legal acts and other agreements
Release policy - transparency
Release policy - user access
Relevance
Completeness

Data revision
Data revision - policy
Data revision - practice
Data revision - studies
Data validation
Data validation – intermediate
Data validation – output
Data validation – source
Dissemination format
Dissemination format - microdata access
Dissemination format - news release
Dissemination format - online database
Dissemination format - other
Relevance - user needs
Relevance - user satisfaction
Reporting agency
Sampling
Source data
Statistical concepts and definitions
Statistical population
Statistical unit
Time period - collection
Timeliness
Timeliness – output
Timeliness – source data
Unit of measure
Valuation

Evaluation

Concerning the accomplishment of the tasks that were planned, according to the "Work Package 2: MCV Ontology, we present an evaluation synthesis:

a) Determine the domain and scope of the ontology – in the present report we address results for a sub-domain;

b) Enumerate important terms of the ontology – within the sub-domain scope we identified about 120 terms, from which 5 are new entries and 7 have review definitions. We identified the context of usage of 73 terms belonging to the core concepts. From this total we are considering other 35 terms to be included on version 2. In general, we consider that most definitions have to be reviewed;

c) Define classes and class hierarchy – we establish hierarchies between several concepts (classes), using "is a" or "kind of" properties;

d) Define properties of classes (slots) and facets of slots – we have defined several properties of classes, namely those for relationships between concepts and properties that designate attributes of concepts. This was formalized in concept maps;

e) Selection of ontology software - we used *IHMC CMapTools* for concept maps and we will use *Protégé* to edit the ontology in OWL.

NEXT STEPS

The working group will proceed with the following tasks:

- finishing the analysis of the sub-domain *data and metadata exchange;*
- revising the correspondent definitions;

- building a first prototype in OWL for concepts related to *data structure definition* and *metadata structure definition* and create some competency questions to evaluate this prototype.

SOFTWARE USED

IHMC CMapTols - http://cmap.ihmc.us/conceptmap.html

REFERENCES

Reference documents on SDMX: Source: http:// www.sdmx.org

A - Standards:

1. Framework for SDMX technical standards
2. SDMX-EDI: syntax and documentation
3. SDMX-ML: schema and documentation
4. SDMX information model: UML conceptual design
5. SDMX implementor's guide
6. SDMX guidelines for the use of web services

B - Guidelines:

1. Content oriented guidelines:
   Annex 1 – Cross domain concepts
   Annex 2 – Cross domain code-lists
   Annex 3 – List of subject matter domains
   Annex 4 – Metadata common vocabulary
2. SDMX user guide

C - Tools:

SDMX Registry

D - Other documents:

GRUBER, T. R. (2009). Ontology. In Ling Liu and M. Tamer Özsu (Eds.) *Encyclopedia of Database Systems.* Springer-Verlag.

GRUBER, T. R. (1993). A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2):199-220. Available online: http://tomgruber.org/writing/ontolingua-kaj-1993.htm

Institute for Human & Machine Cognition (n.d.). Ontology creation for the rest of us… Retrieved August 2010, from http://www.ihmc.us/groups/coe/

OECD GST: Organisation for Economic Co-operation and Development Glossary of Statistical Terms, 2007. Available online: http://stats.oecd.org/glossary/index.htm