Metadata Concepts, Standards, Model and Registries Part B of the Common Metadata Framework (CMF)

Final Report

December 2010, Jana Meliskova

I. Introduction

This final report is addressed to the Conference of European Statisticians Steering Group on Statistical Metadata (METIS). It was prepared by the ad-hoc Task Force, responsible for the preparation of Part B of the Common Metadata Framework (CMF). The aim is to report to the Steering Group about the outcomes of the work on CMF Part B, and to inform about the approach and the activities conducted by the Task Force. Furthermore, it points out the need to ensure a regular maintenance of this product in future in order to keep its content up-to-date. The Steering Group is kindly requested to take the Final Report into the consideration when preparing future activities on METIS.

The Final Report has the following structure:

- chapter II Common Metadata Framework for Statistical Metadata
- chapter III Goal of the CMF- Part B
- chapter IV Ad-hoc Task Force
- chapter V Milestones in the preparation of the CMF-Part B
- chapter VI Recommendations by the Work Session on METIS 2010
- chapter VII Upgrading relationships between resources
- chapter VIII Scope of the CMF-Part B
- chapter IX Maintenance of the CMF Part B
- Annex: Examples of resources descriptions

II. Common Metadata Framework for Statistical Metadata

Part B is an integral part of the Common Metadata Framework for statistical metadata (CMF). The development of CMF was initiated by delegates to the February 2004 meeting of the Joint UNECE-Eurostat-OECD Work Session on Statistical Metadata.

The CMF initiative aims to assist statistical organizations in the adoption, modelling, usage, and implementation of statistical metadata systems and practices across all phases of their statistical business process. This valuable resource is being developed through the collective input of national and international statistical organizations coordinated by the Conference of European Statisticians Steering Group on Statistical Metadata, and the UNECE secretariat.

The CMF is published online via the METIS Wiki (www1.unece.org/stat/platform/display/metis). It is a living document, evolving in line with developments in the field of statistical metadata. It is divided into four parts, each of

which concentrates on different aspects of statistical metadata systems, and provides vital knowledge for anyone working with statistical metadata:

- **Part A - Statistical Metadata in a Corporate Context : A guide for managers**

The material describes the issues surrounding management and governance of statistical metadata system projects. It is focused on the corporate governance of metadata projects. Its target audience is senior managers in statistical organizations. Part A of CMF was positively received and/or evaluated by the Work Session on METIS 2008. Based on its recommendations the Part A was finalized in April 2009. The material is available on the METIS- Wiki.

- **Part B - Metadata Concepts, Standards, Models and Registries**

The material provides information about relevant concepts, international standards and models for statistical metadata. The target audience for Part B is designers and implementers of statistical meta-information systems. The material has been positively received and evaluated by the Work Session on METIS 2010. It was finalized and is available on the METIS- Wiki.

- **Part C - Metadata and the Statistical Business Process**

Information, best practices and other material, to assist metadata developers in statistical organizations to design and develop a statistical information system that is relevant to business requirements. Since the process for statistical surveys is generally the same everywhere, it is possible to build a common business process model for survey work. As a result, the Generic Statistical Business Process Model was developed and finalized in 2009 and is presented in Part C of the CMF: "Metadata and the Statistical Business Process". It is available on the METIS-Wiki.

- **Part D - Implementation**

Focuses on the experiences of national and international statistical organizations that have recently implemented or re-engineered their statistical metadata systems. At the time of writing this report 17 case studies (16 UNECE member countries and UNIDO) have been elaborated. They are regularly updated. The case studies are available on the METIS-Wiki.

III. Goal of the CMF Part B

The issue of standardization of metadata has already been on the agenda of various international groups and organizations for many years (examples include the development of the ISO/IEC 11179 standard, the Data Documentation Initiative (DDI) and the SDMX standard - ISO 17369). Based on requests from UNECE member countries, there has also been an on-going discussion regarding many aspects of international standards for statistics and related metadata within the Work Sessions on METIS.

There is a common understanding in statistical organizations that the use of common standards related to statistics and metadata is indispensable. The number and diversity of existing standards, however, makes it a challenge for statistical experts to understand them and incorporate them efficiently in the statistical production system architecture.

The goal of CMF- Part B is to offer statistical meta-information systems designers an overview of existing resources (standards, concepts, models, best practices and other methodological materials), that are likely to be applicable when designing and implementing a statistical meta-information system. It is designed primarily as an Internet publication, so that it can be kept as up to date as possible.

In general, statistical surveys are conducted in the same way. They follow the same business process, and in fact, the Part C of the CMF is devoted to describing this in the form of the Generic Statistical Business Process Model. From the metadata perspective, this means that a single model for statistical metadata, covering all aspects of the survey life-cycle, is possible.

However, agreement on a single model is very unlikely, and it may not even be practical. What is far more likely is that each program office in each statistical agency will devise its own way of handling metadata. In this case, since metadata are data, too, understanding the metadata for each survey or program will require their own metadata! This replicates the problem, and we aim to avoid this.

Luckily, there is a way around this, through the use of standards. Even though system specifications built by an office for its own use satisfy the needs of that office better than a standard can, there are advantages to using standards over building system specifications locally. The most important are listed below:

1. Standards represent a solution to a business problem that has already been thought through, reviewed, and implemented elsewhere. Time needed to develop a specification is reduced, and systems are built more cheaply.

2. Use of a common specification means that sharing information can be done through the standard rather than with pair-wise agreements. This greatly reduces the burden of interoperability and sharing data or metadata across agencies.

3. Standards are known outside each office that uses them, so tools needed for using a standard may be built by other organizations, systems for implementing a standard may be shared, and knowledge about the use of a standard is readily available.

4. Standards have conformity statements indicating the criteria necessary for claiming that a standard is faithfully implemented. Conformity is a strong claim, and it is usually a sufficient condition for establishing interoperability.

5. Finally, one standard will not fulfil the system requirements for an organization. Even a group of standards may not solve every problem. However, standards are often designed for use with others. As more are used to specify some implementation, the more the savings in development and interoperability costs.

The CMF- Part B addresses these issues.

IV. Ad-hoc Task Force

To make further progress on Part B, an ad-hoc Task Force was convened in early 2009, composed of the following volunteers:

Sérgio Bacelar (INE, Portugal)
Max Booleman (CBS, Netherlands)
Alice Born (Statistics Canada)
Dan Gillman (US Bureau of Labour Statistics)
Hamish James (Statistics New Zealand)
Jana Meliškova (Consultant)
Marco Pellegrino (Eurostat)
Steven Vale (UNECE)

All members of the Task Force are experts in building statistical meta-information and statistical information systems. The Task Force has been operational from May 2009 until October 20010. Its work was organized through monthly teleconferences and the METIS Wiki.

The major goal of the Task Force was to define the scope and content of a knowledge base on standards related to statistical metadata and the document CMF Part B.

When exploring the issue on existing metadata standards, the Task Force benefited from the know-how presented at the Work Sessions on METIS. Furthermore, the knowledge gathered by a large network of experts responsible for the development of metadata standards in international organizations and statistical institutes, has been taken into consideration. All available comments and recommendations to the Part B have been examined when preparing this material.

The preparation of CMF Part B has progressed significantly in 2009-2010. It has now reached the stage where it can be made available to the wide audience of METIS users. As the CMF is a living document, the Part B will never really be finished, but the aim of the Task Force was to reach a stage where the METIS community is confident that the document which has been produced will be a useful resource to the wider statistical community. The Work Session on METIS 2010 confirmed the usefulness of this document. It was recommended to finalize it by the end of 2010.

The activities of the Task Force after the Work Session on METIS 2010 are reaching this deadline. The work of the ad -hoc Task Force for CMF- Part B can be, therefore, considered as accomplished after finalizing this document.

V. Milestones in the preparation of the CMF – Part B

It was agreed that Part B will focus on international standards related to statistical metadata. A valuable source of information for international standards was, of course, the working papers prepared for the METIS Work Sessions 2004-2008. However, because of the voluntary manner of the contributions and, an individual approach to presentations by reporting experts, the information gathered from presented contributions could not give a fully satisfactory overview on standards relevant to statistical metadata to potential users.

In this light, the important goal for the Task Force was to solve many architecture issues, allowing to present an overview in a comprehensive, practical and user friendly form. In the preparation of the CMF- Part B the following milestones should be mentioned:

- **Categorization of Standards**

After the decision that the designers of statistical metadata system are major potential users of the CMF- Part B, the needs of this target group created a basic framework for the presentation of existing standards. The framework contains the following categories:

1. Resources referring to the statistical concepts/constructs;

2. Resources referring to technical specifications for the exchange, storage, documentation and retrieval of statistical data and metadata;

3. Internationally developed models related to statistical metadata;

4. Internationally developed methodological materials and recommendations related to statistical metadata.

The Task Force also thoroughly discussed the role and place of diverse international projects on statistical metadata (e.g. METANET, AMRADS) in the Part B. There was a common understanding that the prevailing goal of such projects was to exchange experiences and knowledge in the building of diverse phases of a statistical meta-information system. None of those resources received a status of international standards, guidelines and/or recommendations. With this in mind, those projects are referenced in a special chapter of the document “Other Resources”.

- **Identification of resources**

Identification of resources was a priority activity for the Task Force. As a result in each of the categories 1, 2 and 3 mentioned above 7 resources have been specified, and 5 resources in category 4. As presented in Figure 1, all together 26 resources have been identified.

There are many other resources that could be described in Part B, for example standards that are only indirectly related to metadata, or national standards and models. The Task Force took a deliberate decision to only focus on the main international resources and standards that are directly related to metadata, but are all the relevant resources described? Should any of those currently included be removed? These questions remain open for further development of Part B.

Furthermore, the Task Force considered the possibility of a second level of resource descriptions. Specific cases for this approach are resources for Statistical Classifications and Statistical Units. An example could be resource descriptions of the main international classifications (ISIC, ISCO, ISCED etc.) Those resources actually refer to a family of resources. Is there a value in developing a second tier of resource descriptions to describe the elements in these resource families? Also this question remains open for further development.

- **Common Terminology**

For the development of CMF-Part B Dublin Core (DC) terminological standard has been used. It was recognized that the DC metadata standard is a simple yet effective element set for describing a wide range of resources as it is the case in the document on Part B. The DC concept predates the Internet. The DC metadata standards have proved to be effective worldwide and are broadly used in electronic publishing. Also, according to

http://semanticweb.org/wiki/Ontology , the DC provides the basis for the most widely referenced ontology on the web, so the DC has served well from “pre web” to “semantic web”. To use DC as a basis for the implementation of Part B has proved to be efficient.

- **Description of Resources**

A common template has been developed for the description of individual resources (see examples in the Annex). An important activity was to map the template to DC terminology. It was agreed that to focus to DC terminology for present version of the Part B. However, a possibility to transfer to other terminology, e.g. Resource Description Framework Schema (RDFS) in the future was taken into the consideration when designing a common template. Task Force members completed template for all identified resources.

- **Relationships between Resources**

The Task Force decided to strengthen the role of links between resources in the document via specific diagrams showing different categories of relationships between resources. This possibility is increasing the potential of the document and its user friendliness. Precondition was that the links in the diagram should be fully consistent with the text in the “Relationships Description to Other Standards” in the template for resource description.

The diagram has been presented for comments at the Work Session on METIS in March 2010 (see Figure 2).

VI. Recommendations of the Work Session on METIS 2010

The Work Session on METIS 2010 considered the progress on the CMF- Part B, gave feedback and supported further work, particularly on the links between resources. The following points were raised during the discussion:

1. The work on the CMF - Part B was seen as useful and important, particularly the description of relationships between standards. It was suggested to add a description of the Portuguese model, as other organizations are planning to adopt this approach.

2. There is a clear need for enlarging and integrating technical standards and to see how the SDMX can be applied within the statistical production process. The application of SDMX along the business process needs to be explored. One possibility could be combining SDMX with other standards (e.g. DDI for micro-data collection).

3. Further integration and expansion of content standards, such as SDMX Content-oriented Guidelines and Eurostat/IMF quality frameworks, seems essential.

4. The Task Force took into the consideration the recommendation (1) when finalizing the document. The recommendations (2) and (3) should be taken into consideration in future activities on this issue.

VII. Upgrading Relationships between Resources

The activities of the Task Force after the Work Session on METIS 2010 concentrated mainly on upgrading the diagram of relationships between resources. It was agreed that the mapping diagram would be produced at 3 levels:

- One diagram showing the relationships between all resources,
- Four diagrams showing the relationships between resources within categories,
- A diagram for each resource, showing the relationships between that resource and other resources.

This upgrade was prepared by Sergio Bacelar, in cooperation with Steve Vale. The software used for this application was IHMC Cmaptools (http://cmap.ihms.us). This software is downloadable from the Internet free of charge.

Using this opportunity, the Task Force would like to express its thanks to Sergio Bacelar (INE Portugal) for his extraordinary effort and contributions to preparing the relationship diagrams.

VIII. Scope of the CMF – Part B

Part B of the CMF is a unique source of information on existing statistical metadata standards. It aims to provide a single point of reference, giving designers and users of statistical meta-information systems basic information about standards related to statistical metadata, as well as links to more detailed materials and resources.

The basic functions of the statistical meta-information system are:

(a) To uniquely and formally define the content and links between statistical objects;
(b) To uniquely and formally describe the content and links between statistical processes; and
(c) To determine all related technical parameters.

These functions are explained in more detail in Part A of the CMF.

To help statistical meta-information system designers decide in which areas of the statistical information system metadata standards should be implemented, the overview of existing statistical metadata standards is presented according to the following groups of standards:

- Statistical concepts;
- Technical standards;
- Models and statistical practices;
- Methodological guidelines and recommendations.

The four groups of standards above should be taken into consideration when designing and implementing a statistical meta-information system. Making links to the Statistical Business Process (see Part C of the CMF), the integration of standards into the statistical meta-information system should be ensured particularly in the following phases of the Generic Statistical Business Process Model (GSBPM):

- Phase 1 - Specify Needs;
- Phase 2 - Design;
- Phase 3 - Build.

It should be pointed out, however, that the work around quality frameworks, discovery metadata etc might call for the need to integrate standards also into other phases of the GSBPM, although the three above mentioned phases are by far the most critical.

The focus in Part B is on the following areas:

(a) Statistical Concepts

This group of metadata standards refers to the content of the statistics. It encompasses internationally accepted statistical standards and/or recommendations that refer to:

- Concepts and definitions used for compiling, disseminating and exchanging statistics;
- Statistical classifications;
- Statistical units;
- Statistical subject matter domains;
- Other standards related to statistical content.

(b) Technical Standards

The metadata standards in this group provide technical specifications for the exchange, storage, documentation and retrieval of statistical data and metadata, as well as other ICT supported activities dealing with the use of metadata for the production of statistics. International standards on Statistical Data and Metadata Exchange (SDMX), metadata registries, Data Documentation Initiative (DDI), Geographical information system (GIS) and other standards are introduced in this chapter.

(c) Models and Statistical Practices

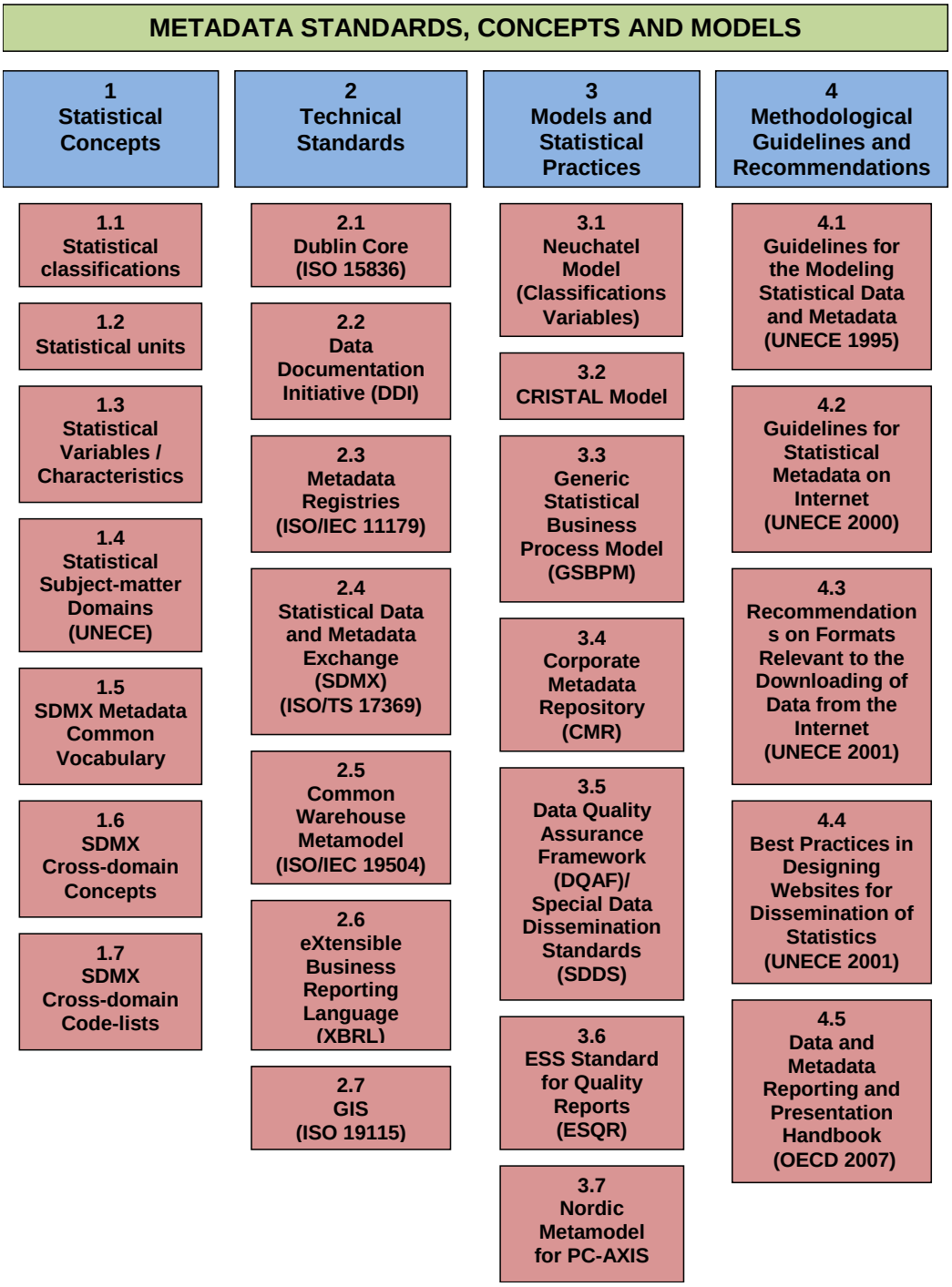
Internationally developed models related to statistical metadata, as well as those developed nationally and recognized and applicable internationally, are presented in this chapter. The Neuchâtel Model on Statistical Classifications and Variables, the Corporate Metadata Repository model, the IMF Data Quality Assurance Framework, and other widely recognized metadata models are presented in this chapter.

(d) Methodological Guidelines and Recommendations

A lot of methodological materials and recommendations related to statistical metadata have been developed in the framework of international cooperation organized by the UNECE together with OECD, Eurostat and other international organizations. Those materials have proved already many times to be an asset for many national and international statistical institutes when building their statistical meta-information systems. "Guidelines for Statistical Metadata on the Internet", and "Best Practices in Designing Websites for Dissemination of Statistics" are examples of such documents. Those and others are introduced in this chapter.

The coverage of these four areas is presented graphically in Figure 1.

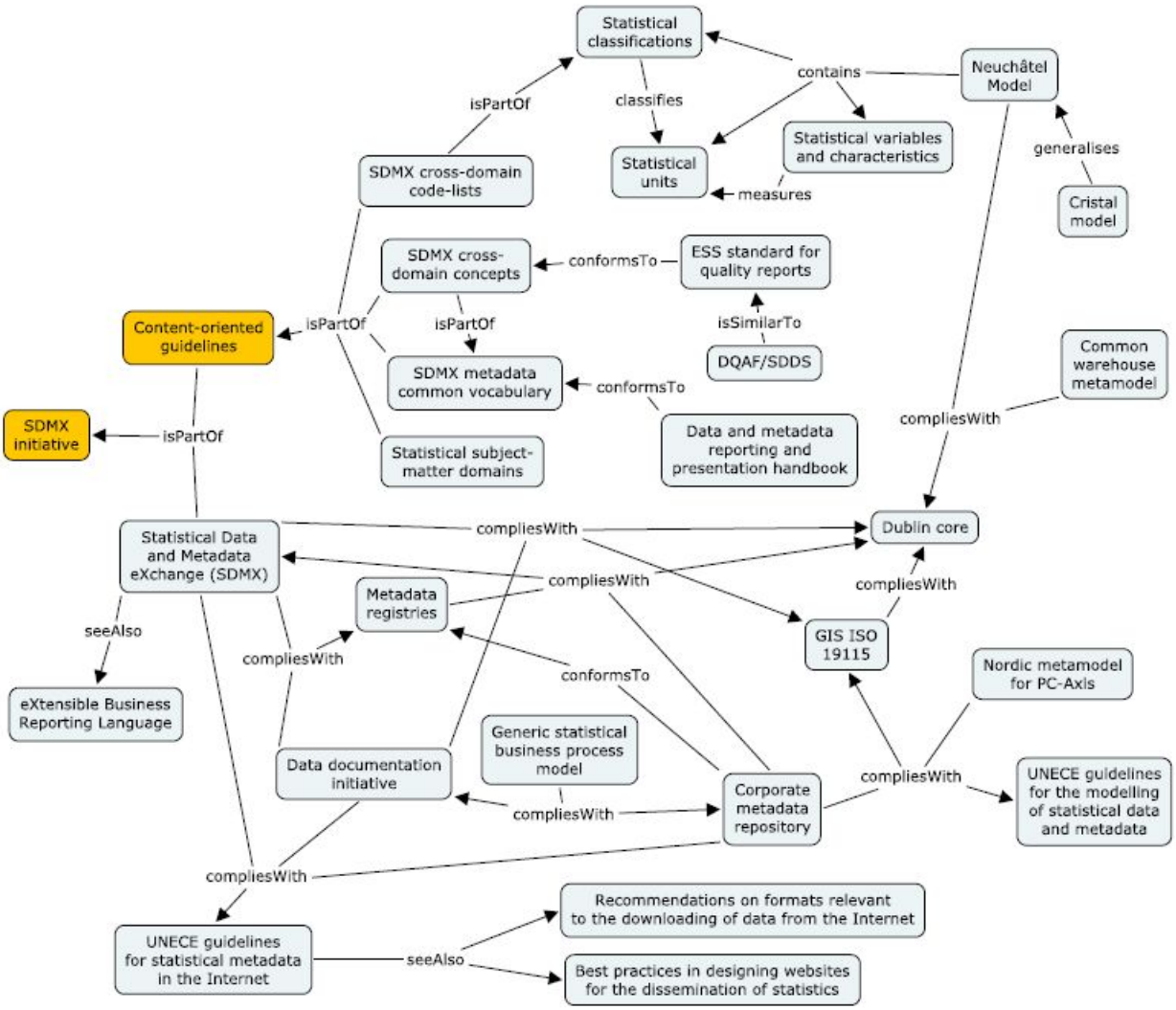
Figure 1: CMF Part B – Metadata Standards, Concepts and Models



Each of the boxes in the above diagram is treated as a resource in Part B, and described with the aid of a common template. Two examples of resource descriptions are provided for illustrative purposes in the annex. The remainder of the resource descriptions can be consulted via the METIS wiki. It is intended that these descriptions are living documents, so any proposals for corrections, changes and additions are very welcome.

Work has also been done to define the nature of the relationships between the resources in this diagram. Figure 2 shows the progress so far.

Figure 2: Relationships between Resources



The links in Figure 2 have been defined as follows, taking into account of existing standards where possible:

classifies – A set of discrete, exhaustive and mutually exclusive observations, which can be assigned to one or more variables to be measured in the collation and/or presentation of data (Metadata Common Vocabulary).

compliesWith – Resource A compliesWith resource B if it is possible to map elements of resource A to elements of resource B. Compliance is not as strict as conformance.

conformsTo – An established standard to which the described resource conforms (Dublin Core)

contains – The described resource includes models for another resource (e.g. “Neuchâtel model” contains a model for “statistical units”).

generalises – Applies the characteristics of a resource to a larger domain and/or includes other functionalities.

instantiates – Represents an abstract concept (like a class in a classification scheme) by a concrete or tangible example (like a code).

IsPartOf – A related resource in which the described resource is physically or logically included (e.g. Dublin Core).

isSimilarTo – Resources are not identical (otherwise the relationship would be **sameAs)** but they show some similarity, for example, they have similar functionalities, and/or the same objective (e.g. the exchange of data and metadata).

Measures – Reading calculating or recording a numerical value (Metadata Common Vocabulary).

seeAlso – Used to indicate a resource that might provide additional information about the described resource (rdfs: **seeAlso**)

IX. Maintenance of the CMF- Part B

The problem still remaining open in the time of writing this report is the maintenance of and update arrangements for the knowledge base, created by the CMF Part B.

Because of quite rapid development in the area of standards related to statistical metadata, the regular update of the CMF Part B is indispensible. Ongoing developments in SDMX, Eurostat/IMF quality frameworks, and integration of DDI and SDMX standards can be some examples of moving forwards in this field.

The Task Force considered some possibilities on how to organize the maintenance of CMF Part B. One proposal was that the Task Force members could take over the responsibility for updating existing resources. Bearing in mind, however, that the Task Force has an ad-hoc status, such an approach could not be operational.

For **maintenance of existing resources**, it is recommended that the UNECE secretariat revises every year the state-of-art in their development and, based on the concrete situation, organizes their update.

For **exploring new resources** and their potential integration into the Part B, it is recommended to use the framework of the Work Sessions on METIS and to consider this issue in the two years interval between Work Sessions.

ANNEX: EXAMPLES OF RESOURCE DESCRIPTIONS

1) SDMX Cross-domain Concepts

Name and version: SDMX Content-Oriented Guidelines, Annex 1: Cross-Domain Concepts (2009 version)

Alternative name: Cross-domain Concepts

Valid: From 2009

Description: Cross-Domain Concepts describe concepts relevant to many statistical domains. The use of these concepts is recommended to promote re-usability and exchange of statistical information and their related metadata between organizations.

Cross-Domain Concepts are part of the SDMX Content-oriented Guidelines and are used in:

- *Data structure definitions*, which define the valid content of data sets;
- *Metadata structure definitions*, which define the valid content of metadata sets;
- Data and metadata messages used for the exchange of data and metadata.

Cross-Domain Concepts have three basic roles:

- As *Dimensions* in a data structure definition, used to identify each statistical observation (for example, a dimension named "Reference Area" would explain which country a specific standard observation refers to);
- As *Attributes* in a data structure definition, qualifying the data further (for example, an attribute named "Unit of Measure" might provide information about whether statistical data are measured in currency units, and if so which currency, or as a pure number);
- As *Attributes* in a metadata structure definition to report metadata about, for example, a data flow, using concepts like timeliness, reference period or data compilation.

Intended use: Any organization providing information about statistical data uses a set of metadata concepts (e.g. frequency of dissemination, reference area, timeliness, type of source data) in order to present the characteristics and quality of the data. Interoperability between data providers will be enhanced when the same concepts are used by many exchange partners and across statistical domains. This is the reason why SDMX recommends the use of this set of common concepts.

Maintenance organization: SDMX consortium

ISO Standard Number: Not applicable

References:

SDMX Cross-Domain Concepts:
http://sdmx.org/wp-content/uploads/2009/01/01_sdmx_cog_annex_1_cdc_2009.pdf

SDMX Content-oriented Guidelines: http://sdmx.org/?page_id=11

Relationships to other standards: **Concept Map**

The concepts included in Cross-Domain Concepts are a subset of the concepts in SDMX Metadata Common Vocabulary.

Language: English

Description last updated / validated: 25 September, 2009

2) Data Documentation Initiative (DDI)

Name and version: Data Documentation Initiative (DDI), version 3.1

Alternative name: [none]

Valid: From October 2009 (version 3.1)

Description: The Data Documentation Initiative is a standard for technical documentation describing social science data. The current version supports description of the full life cycle of a dataset or data collection (see also Generic Statistical Business Process Model).

Intended use: DDI is commonly used as a standard for documenting and describing data for archiving and reuse. It is also suitable for:

Documenting on-going research projects
Documenting secondary uses of data
Creating concept/question/variable libraries
Generating multiple delivery formats for data dissemination or discovery

Maintenance organization: DDI Alliance

ISO Standard Number: not applicable

References:

DDI Help Centre: http://snipurl.com/ddihelp

DDI 3.1 Documentation: http://www.ddialliance.org/specification/ddi3.1

Schema descriptions: http://www.icpsr.umich.edu/DDI/ddi3/Schemas.pdf

Relationships to other standards: **Concept Map**

DDI is expressed as an XML schema

Language: English

Description last updated / validated: 28 December 2010