WP 2
11 February 2010

ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN UNION (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION
AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

Work Session on Statistical Metadata (METIS)
(Geneva, Switzerland, 10-12 March 2010)

PROGRESS ON PART B OF THE COMMON METADATA FRAMEWORK

Paper prepared by the Task Force on Part B,
presented by Jana Meliškova on behalf of the UNECE Secretariat

I. INTRODUCTION

1. The development of a framework for statistical metadata was initiated by delegates to the February 2004 meeting of the Joint UNECE-Eurostat-OECD Work Session on Statistical Metadata (METIS). 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 Common Metadata Framework (CMF) is published online via the METIS Wiki (www.unece.org/stats/cmf), and is 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
  Describes the issues surrounding management and governance of statistical metadata system projects.
- Part B - Metadata Concepts, Standards, Models and Registries
  Clarifies the importance of well-defined, standard terminology. Provides information about relevant concepts, international standards and models.
- 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.
- Part D - Implementation
  Focuses on the experiences of national and international statistical organizations that have recently implemented or re-engineered their statistical metadata systems.

2. 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 Labor Statistics)
Hamish James (Statistics New Zealand)
Jana Meliškova (Consultant)
Marco Pellegrino (Eurostat)
Steven Vale (UNECE)

3. This task force has worked for almost one year through monthly teleconferences and the METIS wiki, to define the scope and contents of Part B, and to start drafting the necessary texts. This paper presents the work accomplished so far, and invites comments on possible additions and improvements. It should be stressed that what is presented at this point in neither final nor complete, and that feedback and assistance in improving Part B is very welcome.

II. BACKGROUND

4. Statistical metadata systems (SMS) should serve statistical organizations as tools for the efficient management and performance of statistical information systems. Globalization has brought several issues in statistical production to greater prominence. There is an increasing user need to make official statistics comparable and easily available both at national and international levels. Clearly, SMS have a lead role in this endeavour. SMS efficiency in this respect is significantly influenced by the standards methods and techniques used.

5. 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. 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 is presented in Part C of the CMF: "Metadata and the Statistical Business Process".

6. While Part A of the CMF, "Statistical Metadata in a Corporate Context: A guide for managers", is focused on the corporate governance of metadata projects and its target audience is senior managers in statistical organizations, the target audience for Part B is SMS designers and experts responsible for SMS implementation.

7. 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 Joint UNECE/Eurostat/OECD Work Sessions on Statistical Metadata.

8. 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 SMS global architecture.

9. **The aim of Part B is therefore to offer SMS designers an overview of existing resources (standards, concepts, models, best practices and other methodological materials), which are likely to be applicable when designing and implementing SMS. It is designed primarily as an Internet publication, so that it can be kept as up to date as possible.**

10. When designing a SMS at the national level, both national and international standards should be taken into consideration. Bearing in mind that national standards often address very specific requirements, this publication is focused on international standards. However, it also includes

information on some of the more important internationally available and/or applicable national models and practices.

III. JUSTIFICATION

11. It is unremarkable to note that every survey and statistical program in each statistical office produces its own data with its own definitions. The consequence of this, however, is remarkable, because this is why we need metadata. One cannot know for sure, *a priori*, what data mean, so metadata describe the data each survey produces, and metadata describe the designs and processes that produce the data. Without metadata, it is not possible to understand and to use data.

12. In general, statistical surveys are conducted in the same way. They follow the same business process, and in fact, 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.

13. 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. First, 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 eliminated, and systems are built more cheaply. Second, 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. Third, 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. Fourth, standards have conformity statements indicating the criteria necessary for claiming a standard is faithfully implemented. Conformity is a strong claim, and it is usually a sufficient condition for establishing interoperability. Finally, one standard will not fulfil the system requirements for an organization. Even a group of standards may not solve every problem. But, 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.

14. This Part B of the CMF addresses these issues.

IV. SCOPE

15. 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 SMS designers and other potential SMS users, basic information about standards related to statistical metadata, as well as links to more detailed materials and resources.

16. The basic functions of the SMS in the statistical 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.

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

18. To help SMS 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.

19. The four groups of standards above should be taken into consideration when designing and implementing an SMS. Making links to the Generic Statistical Business Process (see Part C of the CMF), the integration of standards into the SMS should be ensured in the following phases of the Generic Statistical Business Process Model:

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

20. 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. ISO 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

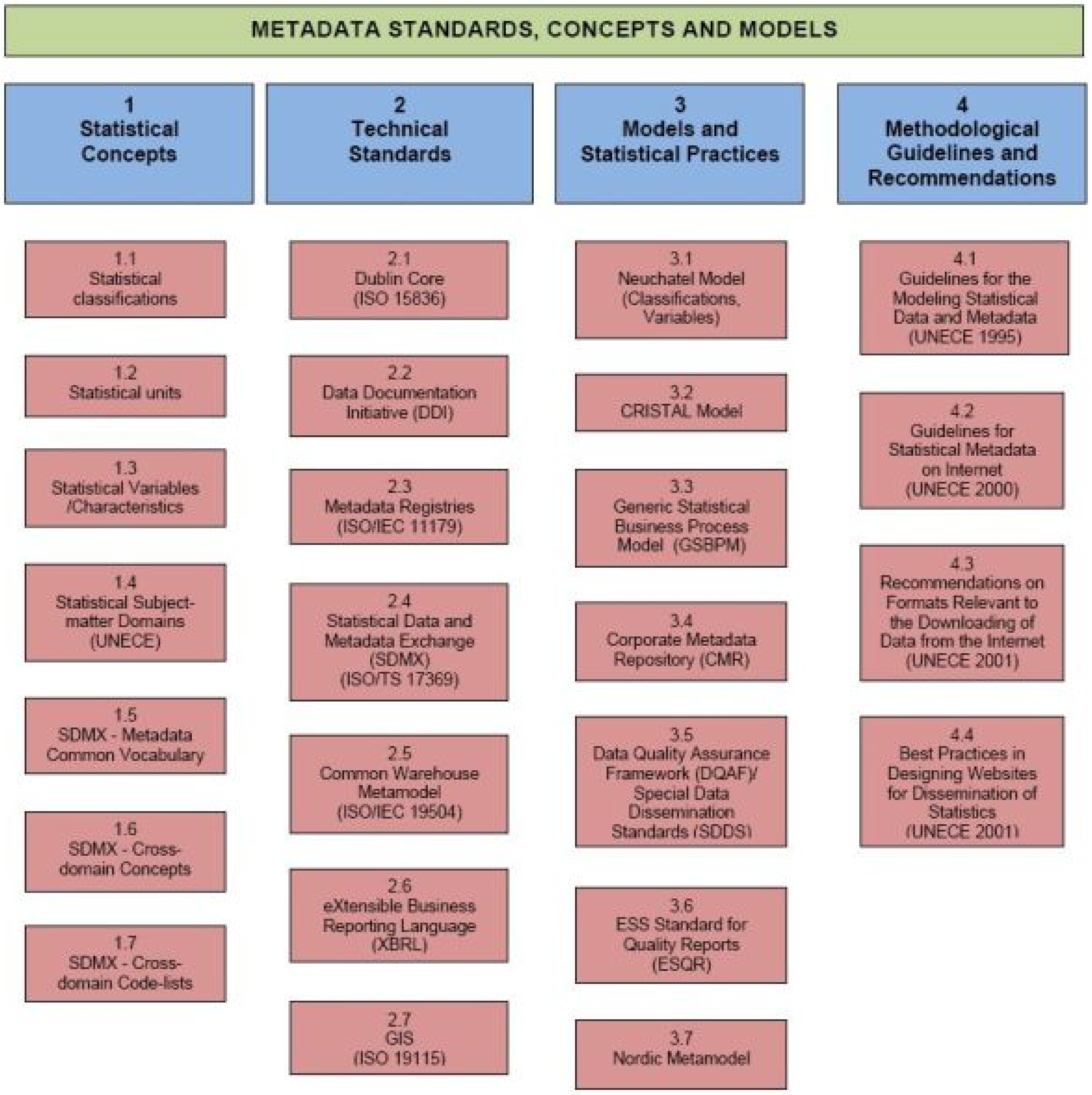
Internationally developed models related to statistical metadata, as well as those developed nationally and recognized and applicable internationally, are presented in this chapter. The Neuchâtel Model on Statistical Classifications and Variables, the Corporate Metadata Repository model, the IMF Data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

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

Figure 1: CMF Part B – Metadata Standards, Concepts and Models (version 8)

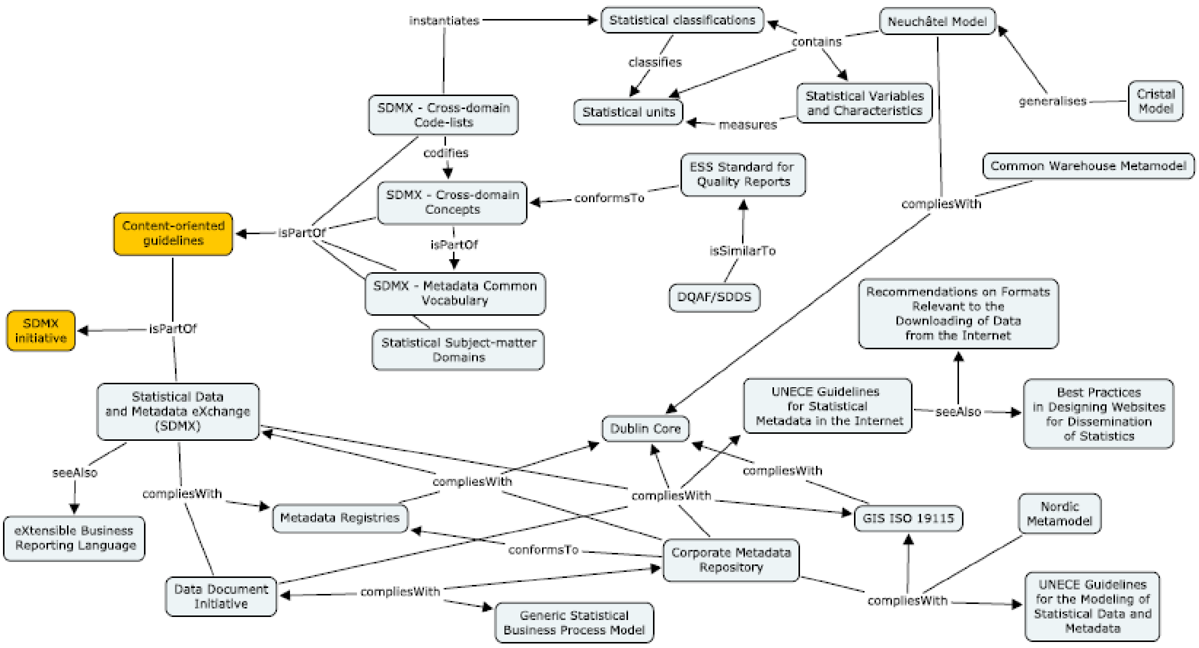

METADATA STANDARDS, CONCEPTS AND MODELS
1 Statistical Concepts
2 Technical Standards
3 Models and Statistical Practices
4 Methodological Guidelines and Recommendations
1.1 Statistical classifications
1.2 Statistical units
1.3 Statistical Variables /Characteristics
1.4 Statistical Subject-matter Domains (UNECE)
1.5 SDMX - Metadata Common Vocabulary
1.6 SDMX - Cross-domain Concepts
1.7 SDMX - Cross-domain Code-lists
2.1 Dublin Core (ISO 15836)
2.2 Data Documentation Initiative (DDI)
2.3 Metadata Registries (ISO/IEC 11179)
2.4 Statistical Data and Metadata Exchange (SDMX) (ISO/TS 17369)
2.5 Common Warehouse Metamodel (ISO/IEC 19504)
2.6 eXtensible Business Reporting Language (XBRL)
2.7 GIS (ISO 19115)
3.1 Neuchatel Model (Classifications, Variables)
3.2 CRISTAL Model
3.3 Generic Statistical Business Process Model (GSBPM)
3.4 Corporate Metadata Repository (CMR)
3.5 Data Quality Assurance Framework (DQAF)/ Special Data Dissemination Standards (SDDS)
3.6 ESS Standard for Quality Reports (ESQR)
3.7 Nordic Metamodel
4.1 Guidelines for the Modeling Statistical Data and Metadata (UNECE 1995)
4.2 Guidelines for Statistical Metadata on Internet (UNECE 2000)
4.3 Recommendations on Formats Relevant to the Downloading of Data from the Internet (UNECE 2001)
4.4 Best Practices in Designing Websites for Dissemination of Statistics (UNECE 2001)

Jana Melišková
15.1.2010

22. 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.

23. Work has also started on defining the nature of the relationships between the resources in this diagram. Figure 2 shows the progress so far.

Figure 2: Relationships between Resources



instantiates
Statistical classifications
contains
Neuchâtel Model
classifies
SDMX - Cross-domain Code-lists
Statistical units
measures
Statistical Variables and Characteristics
generalises
Cristal Model
codifies
ESS Standard for Quality Reports
Common Warehouse Metamodel
SDMX - Cross-domain Concepts
conformsTo
compliesWith
Content-oriented guidelines
isPartOf
isPartOf
isSimilarTo
SDMX - Metadata Common Vocabulary
DQAF/SDDS
Recommendations on Formats Relevant to the Downloading of Data from the Internet
SDMX initiative
isPartOf
Statistical Subject-matter Domains
Statistical Data and Metadata eXchange (SDMX)
UNECE Guidelines for Statistical Metadata in the Internet
seeAlso
Best Practices in Designing Websites for Dissemination of Statistics
Dublin Core
seeAlso
compliesWith
compliesWith
compliesWith
compliesWith
eXtensible Business Reporting Language
Metadata Registries
GIS ISO 19115
Nordic Metamodel
conformsTo
Corporate Metadata Repository
Data Document Initiative
compliesWith
compliesWith
UNECE Guidelines for the Modeling of Statistical Data and Metadata
Generic Statistical Business Process Model

24. The links in Figure 2 have been defined as follows, taking 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).
- codifies - The process of converting verbal or textual information into codes representing classes within a classification scheme, to facilitate data processing, storage or dissemination (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, other domains or includes other functionalities.
- instantiates - To represent 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 (Dublin Core).
- isSimilarTo - Resources are not identical (otherwise the relationship would be sameAs) but they show some similarity, for example, they have similar functionalities, or the same objectives (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).

V. ISSUES

25. The work on Part B is ongoing, and several issues remain to be fully addressed. In addition to providing feedback on the contents of the current version, delegates to the 2010 METIS Work Session are invited to comment on the following:

- Completeness - 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?
- A second level of resource descriptions - Some of the resources described (for example statistical classifications and statistical units) 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? An example could be resource descriptions of the main international classifications (ISIC, ISCO, ISCED etc.).
- Identification of links between resources - Each of the resources described could be said to have links to all of the other resources, however, some links are stronger than others. At present, the diagram in Figure 2 above only attempts to show and describe the stronger links. However, the degree of strength of the links between resources can be seen as a continuum, therefore is Figure 2 sufficiently complete for practical purposes?
- Maintenance and update arrangements - The resource descriptions have been prepared by a small task-force, but they will need updating from time to time. It may therefore help to identify "owners" for each resource description, who will be responsible for maintaining it.

VI. CONCLUSION

26. The work on Part B of the CMF has progressed significantly during the last year. It has now reached the stage where it can be opened to the wider METIS group for comment. As the CMF is a living document, it will never really be finished, but the aim is to reach the stage where the METIS community is confident that it has produced a useful resource, and is happy to promote this to the wider statistical community by the end of 2010. The input of the 2010 METIS Work Session will therefore be extremely useful in helping the Task Force to reach this goal.

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

Format: PDF, SDMX-ML

Language: English

Template last updated / validated: 25 September, 2009

2) Data Documentation Initiative (DDI)

Name and version: Data Documentation Initiative (DDI) version 3.0

Alternative name: [none]

Valid: From April 2008

Description: The Data Documentation Initiative is a standard for technical documentation describing social science data. The current version, DDI 3.0, supports description of the full life cycle of a dataset or data collection (see Generic Statistical Business Process Model).

Intended use: The DDI is commonly used as a standard for documenting and describing data for archiving and reuse. The DDI 3.0 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&nbsp;
- DDI 3.0 Final Schema and Documentation: http://www.ddialliance.org/ddi3/index.html#ddi4
- Schema descriptions: http://www.icpsr.umich.edu/DDI/ddi3/Schemas.pdf

Relationships to other standards: The DDI 3.0 is a development of DDI version 2.3. It is expressed as an XML schema

Format: Not applicable

Language: English

Template last updated / validated: 08 October, 2009