WP.15
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES
(EUROSTAT)**

**ORGANISATION FOR ECONOMIC
COOPERATION AND DEVELOPMENT
(OECD)
STATISTICS DIRECTORATE**

Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS)
(Geneva, 9-11 February 2004)

Topic (iv): Using metadata for searching and finding statistical data in websites and portals

DRAFT METADATA CONTENT STANDARDS FOR STATISTICAL METADATA ON THE INTERNET

Invited Paper

Submitted by OECD[1]

I. INTRODUCTION

1. The need for the formulation of international guidelines and recommendations for the presentation of metadata on the internet has been recognised for some time. More and more, the internet has become the key tool for the dissemination of metadata authored by both national statistical agencies and international organisations. The internet has the potential to make up-to-date metadata readily accessible to a wide range of users with varying degrees of statistical expertise. The technical potential of this dissemination medium with respect to metadata has already been realised, however, as will be discussed below, further work is required in the development of international guidelines and recommendations (and their implementation) on the presentation and content of metadata located on websites before the full capabilities of the internet as a medium for disseminating metadata can be realised, particularly in the context of making international comparisons of national statistical methodologies.

2. This paper states the case for the revision of existing international guidelines and recommendations for the presentation of statistical metadata on the internet for the consideration of METIS. Ultimately, what is envisaged is a single document incorporating existing international metadata content standards, plus any new guidelines and recommendations in this area which METIS believe should be included. An initial outline of what such revised standards should include is provided in Part V of this paper for comment. As can be seen, it would also highlight best practice for statistical website design with respect to the presentation of metadata for searching and finding data in websites, and the interpretation of data. At the moment, there is still considerable variation in both the amount of metadata presented on websites and the form of presentation. The implementation of a minimum set of metadata content standards by national statistical institutes and central banks, etc, for metadata located on the web would facilitate access by international organizations and other users to assess the quality of data, including international comparability.

3. The proposed work will pick up and expand on initial work in this area outlined in the publication, *Guidelines for Statistical Metadata on the Internet*, published by UNSC and UNECE in 2000 (UNSC and

[1] Prepared by Denis WARD, OECD (denis.ward@oecd.org)

UNECE 2000), which is further discussed in the paper prepared by Statistics Norway (Statistics Norway 2004) for this session at this years' METIS. In their paper, Statistics Norway:

- define metadata and the different types of metadata;
- places the provision of metadata firmly in the context of the dimensions of data quality;
- discusses the expanded role of the internet as a key vehicle for the dissemination of statistics and their related metadata; and
- identifies challenges for the future evolution of international metadata standards, including perhaps the need for standards for more "structured documentation" (Statistics Norway 2004, para. 38).

5. The current OECD paper also provides links to other current international initiatives such as the Statistical Data and Metadata Exchange (SDMX) initiative and the review of the IMF's metamodel for the Special Data Dissemination Standard (SDDS), both of which are covered in other sessions of this years' METIS meeting. Finally, the presentation of metadata is also dealt with in the context of a broader range of data presentation issues that will be discussed at the United Nations Statistical Commission (UNSC) meeting in early March 2004. The report submitted for consideration by the Commission outlines proposals for the preparation of a manual containing guidelines and recommended best practice for the presentation of statistical data and metadata disseminated by national agencies and international organisations on various dissemination media. The necessity for such a manual, consolidating existing standards, and developing new recommendations where necessary, stems from the need to further improve data quality (especially interpretability and coherence) and to minimise the burden of reporting data and metadata to international organisations. The report is available on the UNSC website at http://unstats.un.org/unsd/statcom/sc2004.htm.

II. WHAT TYPE OF METADATA?

6. The literature on metadata refers to a number of different metadata classifications describing different types of metadata. The UNSC/UNECE classification identifies three broad types of metadata, namely:

- metadata assisting search and navigation;
- metadata assisting interpretation;
- metadata assisting post-processing.

These have been rearranged somewhat in the Statistics Norway paper which refers to four types of statistical metadata, namely:

- metadata for users of statistical information (for finding and navigation, explaining and post processing);
- quality information;
- process documentation for internal users (for control and improvements); and
- metadata for external data providers (to provide correct data).

I have taken the liberty of rearranging the order of presentation of the four types described by Statistics Norway with regard to "quality information".

7. My preference is towards the Norwegian typology as it makes a useful distinction between metadata required by users to access and interpret statistics, "internal use" metadata and metadata for external data providers. The focus of this paper is the actual content and presentation of the statistical metadata that would most commonly be posted by national agencies and international organisations on their websites and which is designed to both describe the statistics disseminated to users (definitions, classifications, coverage, collection, manipulation, etc) and convey a feel of the quality of the data in relation to a range of expected uses. In the context of the Norwegian metadata classification, this would be the first and second types – metadata for the users of statistical information and quality information.

III. NEED FOR METADATA CONTENT STANDARDS

8. Over the last five years many international and national organisations have placed extensive metadata on the internet, and for many users a search of the website of an agency (or agencies) is the first port of call where there is a need to access metadata to determine the suitability of a given set of statistics for an intended use(s). However, as the Statistics Norway paper points out (in para. 24), there are significant differences between countries when it comes to the organisation and structure of metadata for statistics which are increasingly becoming accessible via on-line dissemination on the web (in html or databases). The evolution of statistical metadata content standards has not really kept pace with IT infrastructure developments. From the perspective of content, there are two broad sets of issues that would need to be covered in any single comprehensive standard for statistical metadata content and presentation, those related to:

- accessibility on the internet. Issues here involve the actual availability of metadata on websites, organisation on the web, provision of search facilities, linkage to data and the financial cost to the user to access the required metadata. Many of these issues are covered in the UNSC/UNECE guidelines;

- significant differences between countries in the actual statistical methodological elements described in metadata located on websites for the same domains. In some instances the problem is merely one of terminology where the same term can have different meanings or different terms can have the same meaning. In other cases, the actual metadata is different. From the viewpoint of an international organisation, where there is a frequent need to compare the practices in use by a number of countries (in the OECD's case, by 30 Member countries) the different metadata content posted on websites makes any meaningful methodological comparisons a time consuming and costly exercise. The need to compare statistics across countries is by no means restricted to users working in international organisations.

IV. WHAT SHOULD A REVISED METADATA CONTENT MANUAL CONTAIN?

9. As mentioned in the Introduction to this paper, what is envisaged is the preparation of a single comprehensive document or manual that would bring together in the one location, key existing and any new international metadata content standards that have been subject to an international review process of some kind. The guidelines would be presented more clearly in the form of a set of specific recommendations than in the case of the earlier UNSC/UNECE document which would be used as a starting point, taking into account developments since 1998 when the document was first authored, in particular, the work of the SDMX initiative and proposed revisions to the IMF's Special Data Dissemination Standard (SDDS) – refer para. 10 below.

10. Before outlining the precise content of any revised manual it is necessary first of all to identify the key issues and recommendations that should be addressed. These would comprise:

- <u>Where</u> metadata should be disseminated by international organisation and national agencies. The key recommendations are that all agencies should:

    - provide access to the metadata required for users to understand the strengths and limitations of the statistics it describes.

    - disseminate such metadata via a range of different media – paper publications, CD-ROMS, etc, however, it is important for all metadata to be available to users on the internet, given that the web provides the most accessible medium for obtaining the most up-to-date metadata. It is also good practice for metadata to be structured in such as way as to meet the needs of a range of users with different needs and/or statistical expertise. In this context a layered presentation of metadata is recommended, progressing from summary metadata to more detailed metadata;

    - keep their metadata up-to-date, incorporating the latest changes in definitions, classifications and methodology, etc;

    - disseminate their metadata free of charge on the web. The OECD strongly supports the notion that metadata describing statistics has a high public good component and should therefore be disseminated free of charge on the internet even if the actual statistics they describe are subject to an organisation's price regime.

- Guidelines to facilitate access to metadata located on websites. National agency and international organisation practices vary significantly with respect to the visibility of metadata located on their websites. In some instances metadata is easily located by users unfamiliar with the site and in others considerable time and effort is required to navigate through the website to obtain the required information, particularly where metadata for a number of different statistical domains are sought. Key recommendations in this area would include:

  - active linkage of metadata to the statistical tables and graphs they describe and vice versa;

  - structuring the metadata for different statistical domains on the basis of some hierarchic classification. Consideration could be given to the adoption of the UN Administrative Co-ordination Committee's (ACC[2]) Classification of Statistics and Statistical Activities as the international standard for metadata. The classification is available at http://unstats.un.org/unsd/methods/statact/acc-class.htm;

  - provision of a local search engine based on free text search;

  - good practice for ensuring the stability of URLs. This is a key issue given the importance of links between websites. The aim would be to develop standards that would minimise broken links;

  - providing the names of contact persons or email addresses where further information may be obtained.

- The methodological items (or metadata elements) that should be incorporated in the metadata posted on websites. Is it possible to identify and obtain agreement at the national and international levels on a minimum or core set of metadata items that would be relevant to all statistical domains? This issue is at the heart of current problems and difficulties of comparing methodologies used by different countries in the compilation of the statistics they disseminate. The notion of a minimum core set of metadata required for the correct interpretation of statistics has been discussed at previous meetings of METIS and indeed such a list is included in the UNSC/UNECE guidelines (UNSC/UNECE 2000, p. 5). Similarly, the more comprehensive and hierarchic metamodel for the IMF's SDDS, which is currently being revised, provides another such core list. METIS could consider the possibilities of using the SDDS metadata as a generic model.

- A set of practices to be followed by international organisations to minimise the metadata reporting burden of national agencies? In addition to a perceived lack of co-ordination between international agencies, national agencies faced with the burden of providing metadata to different international agencies, often comment on their use of different metadata templates for the same statistical domains. They also comment on how much easier life would be if different international agencies used the same metamodel (or at least a common core template) so that one set of metadata compiled by the national agency would meet the needs of many/all/most international agencies. Another form of co-ordination would involve the linkage of metadata held on various national and international repositories in lieu of direct collection and/or duplicate storage on different databases.

- The need to adopt a common set of terminology. Considerable resources are often expended by international organizations in verifying text, etc, to ensure that methodological descriptions are as consistent as possible between countries. Not only does the process of metadata verification entail a duplication of effort but it also results in dissemination of different methodological terminology, especially where translation of methodological text into another language is necessary. Ideally, methodological descriptions of the same national statistical collections published by different international agencies should be identical with regards terminology. A mechanism for achieving this would be the rigorous use of terminology imbedded in the various international statistical guidelines and recommendations. This could be facilitated by the use of glossaries published by international organizations which contain definitions

[2] Now known as the Committee for the Co-ordination of Statistical Activities (CCSA) which is a body of representatives from all UN and non-UN international organisations involved in statistical activity. The CCSA normally meets once a year.

derived from these standards. Examples of such glossaries are those maintained by the OECD, Eurostat and UNSD[3].

The Metadata Common Vocabulary (MCV) developed by Eurostat and the OECD under the umbrella of the SDMX initiative is specifically aimed at identifying commonly used terms to describe the different types of metadata (SDMX 2003). It is intended to be used by international organizations and national statistical agencies. The MCV contains a core set of metadata items and their related definitions and is designed to improve the standardization of metadata content for the purposes of data exchange and to promote the use of common nomenclatures that can foster international comparability of international data. The current version of the MCV (available on the SDMX website at www.sdmx.org) contains several fields – term, definition, source, URL to definition source where available, related terms and context.

11 The introduction to the proposed revised metadata manual would reinforce the essential need for the provision of metadata, define metadata and the different types of metadata and users of metadata. Much of this information is already located in the UNSC/UNECE guidelines. The subsequent sections of the manual would also describe good practice for each of the key issues outlined in the previous paragraph.

V. DRAFT OUTLINE

12. Flowing from the above discussion, the content of the revised metadata manual would therefore comprise:

	Source(s) of text
A. Introduction • Essential need for metadata in the context of data quality dimensions • Definition of metadata • Outline of different types of metadata • Needs of different types of metadata users • Overview of different metadata dissemination media • Aims of manual – metadata issues covered (and not covered) • Outline of manual	Much of this information/text would come from the UNSC/UNECE 2000 - updated as required.
B. Guidelines on where metadata should be disseminated by national agencies and international organisations • In paper publications • In electronic dissemination media – CD-ROMS – databases – internet • Free or priced access to metadata disseminated on the internet	New text required.
C. Guidelines on presentation of metadata on the internet to facilitate access • Active linkage of metadata to the statistics they describe and vice versa. • Hierarchic statistical domain (subject matter) classification • Local search engine based on free text search • Good practice for ensuring stability of URLs • Provision of contact persons or email addresses	Based on UNSC/UNECE 2000 guidelines - updated as required. Domain hierarchy based on ACC Classification of Statistics and Statistical Activities.

[3] Refer to the OECD Glossary of Statistical Terms (OECD 2002); Eurostat's CODED (Eurostat 2003) and the UNSD glossary (UNSD 2002)

D. Minimum core metadata items to be provided to assist interpretation of statistics by users.	Revised Metamodel for IMF’s Special Data Dissemination Standard (SDDS)
E. Practices to be followed by international organisations to minimise metadata reporting burden of national agencies.	UN Fundamental principles of Official Statistics Additional text reqired.
F. Adoption of common metadata vocabulary or nomenclature	SDMX Common Metadata Vocabulary Comprehensive glossaries published by UNSD, Eurostat, OECD

G. Key references

- Links to existing international metadata standards and glossaries

- Selected examples of good national and international practice

VI. PROCESS FOR REFORMULATION AND ELABORATION OF METADATA STANDARDS

13. Another issue that should be considered by METIS is the process for the adoption of a draft revised document or manual as a comprehensive new international standard. Issues to be considered include the:

- role of METIS in preparing an initial version of the revised manual and the identification of a process for obtaining detailed input from national delegates and the delegates from other international organisations that attend METIS following the February 2004 meeting;

- preparation of draft document by the end of 2004 for consideration by UNSC (in 2005) and the process for incorporating input from a wider group of countries.

VII ISSUES FOR CONSIDERATION/DISCUSSION BY METIS IN FEBRUARY 2004

14. In summary, the specific issues for consideration/discussion by METIS in February 2004 are:

a) Is there a need for the revision and further elaboration of the international metadata standards contained in the UNSC/UNECE document published in 2000 that would incorporate the more recent developments outlined above and during other METIS sessions this year?

b) If there is, what additional aspects/issues should be included in a revised, perhaps more comprehensive document or manual?

c) Finally, what would be the process for obtaining detailed input from METIS delegates, a wider range of countries and for endorsing the draft proposals by METIS and subsequently, the UNSC?

REFERENCES

Eurostat, 2003, Eurostat concept and definitions database (CODED), Eurostat, Luxembourg, available at http://forum.europa.eu.int/irc/dsis/coded/info/data/coded/en.htm [Accessed 29 July 2003]

OECD, 2002, *Glossary of Statistical Terms*, OECD, Paris, available at http://cs3-hq.oecd.org/scripts/stats/glossary/index.htm [Accessed 29 July 2003]

Statistical Data and Metadata Exchange (SDMX) – BIS, ECB, European Community, IBRD, IMF and OECD, 2003, Metadata Common Vocabulary, available at www.sdmx.org [Accessed 15 December 2003]

Statistics Norway, 2004, "Statistical Metadata on the Internet Revisited". Invited paper to the Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS), Geneva, 9-11 February 2004.

UNSC and UNECE, 2000, *Guidelines for Statistical Metadata on the Internet*, Conference of European Statisticians Statistical Standards and Studies – No. 52, United Nations, Geneva Available at http://www.unece.org/stats/documents/statistical_standards_&_studies/52.e.pdf [Accessed 16 December 2003]

UNSC and UNECE, 2000a, *Terminology on Statistical Metadata*, Conference of European Statisticians Statistical Standards and Studies – No. 53, United Nations, Geneva Available at http://www.unece.org/stats/documents/statistical_standards_&_studies/53.e.pdf [Accessed 16 December 2003]

UNSD, 2002, Definitions for United Nations Common Database, UNSD, New York, available from http://unstats.un.org/unsd/cdb/cdb_help/cdb_quick_start.asp [Accessed 29 July 2003]