

# OECD WORKSHOP ON BUSINESS AND CONSUMER TENDENCY SURVEYS, WARSAW – 14 SEPTEMBER 2004

## Session 2: DRAFT BTS/COS METADATA STANDARDS

### a. Introduction

1. In recent years greater emphasis has been given to the importance of ensuring that statistics published by national statistical institutes, international organisations and other agencies are accompanied by adequate methodological information or metadata<sup>1</sup>. Many statistical agencies have embodied their corporate policy on the provision of metadata in their dissemination standards and author guides. The need for such methodological information arises from a desire to lend transparency to the data so that the typical end-user can make an informed assessment of their usefulness and relevance to his or her purpose.

2. There are significant differences between statistical agencies when it comes to the organisation and structure of metadata for statistics which are increasingly becoming accessible via a wider range of dissemination media, in particular, on-line dissemination on the web (in html or databases). The evolution of statistical metadata content standards has not really kept pace with ITC infrastructure developments. From the perspective of content standards, there are two broad sets of issues:

- accessibility of the metadata. In the context of internet dissemination, issues here involve the actual availability of metadata on websites, linkage to data and the financial cost to the user to access the required metadata; and
- significant differences between countries and institutes in both the amount and actual content of the metadata provided for data disseminated. In some instances the problem is merely one of terminology, whereas in others, the metadata is different. From the viewpoint of an international organisation, where there is a frequent need to compare practices used by a number of countries, different metadata content posted on websites or published elsewhere 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.

3. This paper outlines draft metadata standards for business tendency and consumer opinion survey data for consideration and discussion at this workshop. These standards comprise three elements:

- recommended good practice for the dissemination of metadata;
- recommendations for standard terminology for metadata preparation; and
- a set of common metadata items.

4. The draft standards presented align with existing standards where they exist or with standards currently being developed. In the main, these standards have been (are being) developed for quantitative statistics though in almost all instances they are also relevant for BTS / COS data. The current workshop provides a useful opportunity to identify and insert any aspects specific to qualitative statistics into metadata standards currently being developed.

---

<sup>1</sup> The International Standards Organisation (ISO) definition of metadata is “data that defines and describes other data”. Metadata in the context of this paper is more akin to the term statistical metadata defined by the United Nations Statistical Commission (UNSC) as “..... information on data – and about processes of producing and using data. Metadata describe statistical data and – to some extent – processes and tools involved in the production and usage of statistical data.”.

## **b. Recommended good practice for the dissemination of metadata**

5. This Section outlines key elements of good practice for the dissemination of metadata which specifically relate to where metadata should be disseminated by international organisations and national agencies and the provision of access to metadata. Issues relating to the use of a common set of terminology in the preparation of metadata and to the methodological items (or metadata elements) that should be incorporated in metadata disseminated are discussed below in Sections (c) and (d).

6. Recommended good practice for the dissemination of metadata requires that all statistical agencies should:

- compile metadata that will enable users to understand the strengths and limitations of the statistics it describes and to assess the relevance of the data to their particular need(s);
- ensure that users have ready access to such metadata through its dissemination 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 a way as to meet the needs of a range of users with different needs and/or statistical expertise;
- 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. There is strong support for 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 and paper publication versions, etc, are subject to an organisation's price regime;
- adopt good practice for ensuring either the stability of URLs (Uniform Resource Locators) or providing links between the old and new URLs that will redirect users to the new address. This is a key issue given the importance of links between websites<sup>2</sup>.
- provide contact details, email address, etc, where further information about concepts, definitions and statistical methodologies may be obtained.

## **c. Adoption of a common set of terminology for metadata preparation**

7. As mentioned in para. 2 above, differences between countries and institutes in both the amount and content of metadata disseminated complicates international data comparisons. 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. Ideally, a set of consistent terminology should be used by different institutes to describe the same concept. To facilitate this, international organizations such as the OECD, Eurostat and the United Nations Statistical Division have developed and published extensive statistical glossaries containing definitions relevant to all types of statistical data, including BTS/COS. An example is the glossary published at the back of the OECD publication, *Business Tendency Surveys: A Handbook*<sup>3</sup>.

8. A more recent example of a glossary relevant for the compilation of metadata is the Metadata Common Vocabulary (MCV) developed by Eurostat and the OECD under the umbrella of the Statistical

---

<sup>2</sup> The World Wide Web Consortium (W3C) document “Cool URIs don’t change” (available at <http://www.w3.org/Provider/Style/URI>) outlines the case for maintaining stable URLs and best practice for designing URLs.

<sup>3</sup> The definitions in the Handbook glossary are also available in the *OECD Glossary of Statistical Terms* available at <http://cs3-hq.oecd.org/scripts/stats/glossary/index.htm>. The Handbook itself is available at <http://www.oecd.org/dataoecd/29/61/31837055.pdf>

Data and Metadata Exchange (SDMX) initiative<sup>4</sup>. It is specifically aimed at identifying commonly used terms to describe the different types of metadata, and 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 promote the use of common terminology and improve the standardization of metadata content for the purposes of data exchange and international comparisons. The current version of the MCV (available on the SDMX website at [www.sdmx.org](http://www.sdmx.org)) contains several fields – term, definition, source, URL to definition source where available, related terms and context.

9. It is recommended that the definitions provided in the glossary of the OECD publication, *Business Tendency Surveys: A Handbook* and in the MCV be adopted as a basis for common nomenclature in the preparation of metadata for BTS / COS.

#### **d. Adoption of a set of common metadata items**

10. While it is obvious that all BTS / COS data disseminated by national agencies and international organizations should be accompanied by appropriate metadata, it is less obvious as to precisely what the “appropriate” metadata should be. In order to achieve some level of conformity in the metadata to be disseminated by different institutes, it is necessary to identify some *common metadata items*. In this way, users will have more certainty about the actual metadata they can expect to be able to find across countries which will also facilitate comparisons of national practice.

11. The draft list of common metadata items outlined below in the Annex to this paper should be seen as a kind of structured questionnaire, where the 34 individual “questions” or metadata items to be filled in, presented in column (3), are grouped under six broad headings, presented in column (1). The list of common metadata items specified in the Annex are intended to be general in the sense that they should have a good chance of being relevant to the whole range of BTS / COS data. In order to end up with a relatively simple and manageable list of metadata items, the ambition is to be able to place around 80% of the metadata in discrete metadata items or fields. Provision will be made to place the remaining 20% of metadata into “other items”, “other aspects”.

12. The items are arranged in the list of common metadata items in two levels (top level and child level). Actual metadata text is always provided at the child level. The function of the top level is primarily to group child level items to facilitate user access. A further elaboration of the list of common metadata items could be the drawing of a clear distinction between metadata that is primarily “descriptive”, i.e. actual geographic coverage, sources of weights, etc., and metadata that could be regarded as qualitative in nature such as departures from international guidelines, recommendations and classifications, breaks in time series, etc. The draft list allows for the provision of such information across the whole range of child level items.

13. The application of the headings with respect to the actual inclusion of metadata by different institutes is obviously not mandatory in that all child level items do not have to be filled in. The number of items to be populated with metadata text depends on the metadata objectives of each statistical agency or institute and its resource capacity to maintain the metadata. However, to facilitate international comparisons it is essential to include under a heading, and at the appropriate level of detail, all available metadata matching that heading. A definition of each heading is provided in the SDMX Metadata Common Vocabulary (MCV). In this way, metadata will be much easier to locate and comparable in terms of content, and thus more useful. It will also enable the mapping of metadata maintained by other international organizations and national agencies and facilitate the exchange of metadata.

---

<sup>4</sup> SDMX is a consortium of seven international organisations (Bank for International Settlements (BIS), European Central Bank (ECB), Eurostat, IMF, OECD, United Nations Statistical Division (UNSD) and the World Bank working together to develop standards for the exchange of data and metadata. Information about SDMX is available at [www.sdmx.org](http://www.sdmx.org)

14. The draft list of common metadata items is aligned with similar lists and standards being developed by several international within the SDMX initiative for final agreement in 2005. This is especially important because the SDMX standards being developed are intended to be set in accordance with what national statistical organisations can be expected to provide by linking or mapping from their own metadata systems. Furthermore, a definition for each of the common metadata items will be developed and maintained in the Metadata Common Vocabulary (MCV) being developed within the SDMX initiative. The draft list in the Annex is also consistent with the list of metadata items listed in the OECD publication, *Business Tendency Surveys: A Handbook* (refer pp. 49-51).

#### **e. Discussion issues**

15. Feedback / comment are especially sought with respect to:

- the recommended good practice for the dissemination of metadata outlined in para. 6;
- the relevance of the BTS / COS terminology and definitions provided in the OECD publication, *Business Tendency Surveys: A Handbook* (refer footnote 3 for URL) and the metadata common vocabulary (MCV) (available at [www.sdmx.org](http://www.sdmx.org) ). Are any additional definitions required?
- the relevance of the draft list of common metadata items provided in the paper Annex. What significant changes are required to make the list more relevant to BTS / COS data?

## Common metadata items (draft version – 24 August 2004)

(1) Top level <sup>1</sup>	(2) Description derived from the Metadata common vocabulary (MCV)	(3) Child level <sup>1</sup>	(4) Description of metadata requirement (derived from the Metadata common vocabulary (or “non-MCV” description where no MCV definition currently exists)	(5) Notes on possible contents <sup>2</sup>
Source organisation	Organisation from where data was submitted / extracted	<u>Contact person and organisation</u>	Contact person, title, unit, organisation, phone number, fax, number, email, city, country, postal code [non-MCV]	
Data characteristics and collection		Unit of measure used	Refers to the unit in which associated values are measured, e.g. USD [non-MCV]	
		Power code	Power of 10 by which the reported statistics should be multiplied, e.g. “6” indicating millions of USD. [non-MCV]	Natural numbers
		<u>Data source(s) used</u>	List <u>data source(s)</u> used (administrative data, household survey, enterprise/establishment survey, etc).	
		<u>Name of collection / source used</u>	Refers to full title of survey collection, administrative source, database or publication from where the data were obtained. [non-MCV]	
		<u>Variables collected</u>	List of variables collected or provision of questionnaire [non-MCV]	
		<u>Sampling</u>	Refers to information on <u>sample size</u> , <u>sample frame</u> , <u>sample updating</u> , <u>sampling error</u> , <u>sample (other)</u>	
		<u>Reporting unit</u>	The unit for which data are collected	
		<u>Periodicity</u>	The time distance between observations (whether stock or flow). Values: Yearly, Quarterly, monthly, irregular <other?>	
		<u>Reference period</u>	Period of time the data refer to. For business tendency or consumer opinion surveys this field could also refer to the forecasting horizon. [non-MCV].	
		Base period	The period of time for which data used as the base of an index number, constant prices data or other ratio, have been collected.	
		Date last input	Refers to the date on which the data was last received from	

Top level <sup>1</sup>	Description derived from the Metadata common vocabulary (MCV)	Child level <sup>1</sup>	Description of metadata requirement (derived from the Metadata common vocabulary (or “non-MCV” description where no MCV definition currently exists)	Notes on possible contents <sup>2</sup>
(1)	(2)	(3)	(4)	(5)
		received from source	the source, e.g. national agency or international organisation. [non-MCV]	dd/mm/yyyy
		Non-response	Indication of the level of non-response, together with information on procedures to deal with non-response (e.g. substitution, imputation, etc)	
		Link to <u>Release calendar</u>	Refers to a link to a general statement on the schedule of release of data.	
		Other Data characteristics and collection		
Statistical population and <u>scope of the data</u>	The scope is the coverage or sphere of what is to be observed. It is the total membership or population of a defined set of people, object or events.	<u>Statistical population</u>	Target population (the statistical universe about which information is sought).	Departures from international guidelines and recommendations
		<u>Geographic coverage</u>	The geographic area covered by the data. [non-MCV]	Information on exceptions and departures from international guidelines and exceptions
		<u>Sector coverage</u>	The range of sectors covered by the data [non-MCV]	Information on exceptions and departures from international guidelines and exceptions
		<u>Institutional coverage</u>	The range of institutions covered by the data [non-MCV]	Information on exceptions and departures from international guidelines and exceptions

Top level <sup>1</sup> <b>(1)</b>	Description derived from the Metadata common vocabulary (MCV) <b>(2)</b>	Child level <sup>1</sup> <b>(3)</b>	Description of metadata requirement (derived from the Metadata common vocabulary (or “non-MCV” description where no MCV definition currently exists) <b>(4)</b>	Notes on possible contents <sup>2</sup> <b>(5)</b>
		<u>Item coverage</u>	The range of items covered by the data [non-MCV]	Information on exceptions and departures from international guidelines and exceptions
		<u>Population coverage</u>	The population covered by the data [non-MCV]	Information on exceptions and departures from international guidelines and exceptions
		<u>Product coverage</u>	The range of products covered by the data [non-MCV]	Information on exceptions and departures from international guidelines and exceptions
		Other coverage	Other issues and information concerning the coverage of the data [non-MCV]	Information on exceptions and departures from international guidelines and exceptions
<u>Statistical concepts and classifications used</u>		<u>Key statistical concepts used</u>	A statistical concept is a statistical characteristic of a time series or an observation. This item should define key statistical concepts included in the domain of study	Departures from concepts defined in international guidelines and recommendations
		<u>Classification(s) used</u>	A classification is 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. This item should list the name of all classifications actually used in the compilation of the data.	Departures from international classifications
<u>Manipulation</u>		<u>Aggregation &amp;</u>	Aggregation is the combination of related categories, usually	

Top level <sup>1</sup>  <b>(1)</b>	Description derived from the Metadata common vocabulary (MCV)  <b>(2)</b>	Child level <sup>1</sup>  <b>(3)</b>	Description of metadata requirement (derived from the Metadata common vocabulary (or “non-MCV” description where no MCV definition currently exists)  <b>(4)</b>	Notes on possible contents <sup>2</sup>  <b>(5)</b>
and dissemination		<u>consolidation</u>	within a common branch of a hierarchy, to provide information at a broader level to that at which detailed observations are taken	
		<u>Estimation</u>	Estimation is concerned with inference about the numerical value of unknown population values from incomplete data such as a sample.	
		<u>Imputation</u>	Refers to procedures for entering a value for a specific data item where the response is missing or unusable.	
		<u>Validation</u>	A procedure which provides, by reference to independent sources, evidence that an enquiry is free from bias or otherwise conforms to its declared purpose. It may be applied to a sample investigation with the object of showing that the sample is reasonably representative of the population and that the information collected is accurate. Refers to processes applied for the <u>verification of data, data confrontation, and data reconciliation</u>	
		<u>Index type</u>	<u>Index type,</u>	
		<u>Weights</u>	Refers to information on <u>sources of weights, nature of weights, period of current index weights, frequency of weight updates, weights (other)</u>	
		<u>Seasonal adjustment</u>	Seasonal adjustment is a statistical technique to remove the effects of seasonal calendar influences operating on a series. Seasonal effects usually reflect the influence of the seasons themselves either directly or through production series related to them, or social conventions. Should provide information to enable users to make an assessment of the validity of the seasonal adjustment applied. Such information would comprise: a short description of the method (software) used; the main parameters of the adjustment (e.g. additive v. multiplicative decomposition) and some of the derived information (e.g. trading-day weights). [non-MCV]	

Top level <sup>1</sup>	Description derived from the Metadata common vocabulary (MCV)	Child level <sup>1</sup>	Description of metadata requirement (derived from the Metadata common vocabulary (or “non-MCV” description where no MCV definition currently exists)	Notes on possible contents <sup>2</sup>
(1)	(2)	(3)	(4)	(5)
		Other manipulation & adjustments	Manipulation and adjustments not mentioned under the headings Aggregation & consolidation, Estimation, Imputation, Validation, Index type, Weights, Sampling, Seasonal adjustment	
		<u>Dissemination format(s)</u>	Refers to the different dissemination media used to disseminate the data, e.g. news release, paper publication, on-line or database, CD-ROM or other. [non-MCV]	
Other aspects		<u>Quality comments</u>	Gives the possibility for data managers to insert comments of quality aspects or general evaluation of quality, as seen from a user perspective.	
		<u>Other comments</u>	Other important aspects	

1. Top and Child level headings underlined equate with those currently proposed for use by IMF and Eurostat in their metadata dissemination models.
2. Allows insertion of information on departures from existing international guidelines and recommendations for any Child level metadata items as well as information on series breaks.

