



Session 5: Tools and practices for collecting quality metadata from data providers and informing users on the quality of statistics

The use of SDMX standards for retrieving metadata from different providers and achieving rapid dissemination: issues for discussion

Prepared by Marco Pellegrino – Eurostat¹

I. INTRODUCTION

1. The objective of this paper is to highlight a number of areas where work is currently taking place within Eurostat and the other SDMX sponsoring organisations² to facilitate the standardisation of metadata concepts used and for establishing a stronger coordination of metadata requirements by international organisations, so that comparable data and metadata can be made available by the relevant providers more easily, reducing redundancies and minimising reporting effort. These tasks entail development of metadata standards in both the IT area (such as for the use of common XML formats and tools) and in “metadata content”.
2. The implementation across Eurostat, in 2004, of a standardised format for releasing information on the data disseminated over the Internet has been a success. External users can find the information in a standard template built on the concepts used in the SDDS format (data coverage, geographic coverage, periodicity, timeliness, release calendar, terms and conditions, summary methodology, etc) slightly adapted to fit Eurostat needs and purposes. At present, more than 560 reference metadata files are posted on the web site and linked to the data tree at various levels (domain, sub-domain, statistical table). Most files (around 70% of the total) have been updated at least once during the last year. The visibility of these metadata is very high and statistics on web usage show an average of more than 3 thousand consultations per day. A policy for monitoring the quality of metadata descriptions is in place, together with training initiatives and specific actions for increasing the coverage of Eurostat sub-domains and tables.
3. In spite of this success, users are still confronted with a lot of problems when trying to retrieve comparable methodological information from different countries and providers of the European Statistical System. The ESS as a whole is not yet able to provide users with a common set of standardised, comparable and re-usable metadata describing both European and national statistics. The adoption of a standard format at European level should foster comparability of metadata; but on the other hand our first self-assessment acknowledged that most metadata files are weak in bringing to the reader information on some methodological items which are managed at country level.

¹ Marco Pellegrino (Eurostat, Directorate B: "Statistical methods and tools; dissemination", marco.pellegrino@cec.eu.int). The views expressed in this paper are those of the author and should not be attributed to Eurostat management.

² The SDMX sponsoring organisations are: Bank for International Settlements (BIS), European Central Bank (ECB), Eurostat, International Monetary Fund (IMF), OECD, United Nations Statistical Division (UNSD) and World Bank.

II. USING SDMX STANDARDS FOR METADATA EXCHANGE AND SHARING

4. In some important domains (for instance, HICP and the so-called Principal European Economic Indicators, PEEI) ideas and plans for improving the exchange and dissemination of specific methodological elements at country level have already been discussed within the relevant working groups. The technical solution is being developed through the implementation of a common platform for dissemination based on SDMX standards (SODI) which provides a new framework for making available not only data but also metadata of good quality within EU, in a standard format, for the points where this can offer a value added to the user.
5. Recent developments in the use of SDMX technical standards (ISO/TS/17369) encourage the specification of formal rules for formatting data and metadata, so that these can be exchanged, read and processed by computers without manual intervention. This involves the use of well-specified metadata concepts that are considered to be cross-domain and of universal use, such as “periodicity”, “timeliness”, “data source”, “statistical adjustment” or “compilation”, as well as those that are specific to a statistical subject-matter domain. A web-service, using information about web locations of data and metadata, can then navigate, find and automatically process the information for analytical and dissemination purposes. The progress made on the identification of commonalities in the existing metadata repositories and on the standardisation of the terminology used (linking to the Metadata Common Vocabulary – MCV – elaborated within SDMX) will help reducing the metadata reporting burden of national institutes and, at the same time, will improve the quality and consistency of metadata descriptions across countries.
6. The new version 2 of the SDMX set of standards represents a major advance over version 1, facilitating the exchange of statistical metadata through the use of web services and mark-up languages. It supports richer and more complex data/metadata structures and it allows querying metadata across various sites for retrieving a customised reporting.
7. The core of the SDMX work on metadata is the concept of “Metadata Structure Definition”, which is the equivalent of the “Data Structure Definition” (also known as “key family”) for metadata. The Metadata Structure Definition defines the content of metadata and identifies the structures to which metadata can be attached (metadata are “referenced” from the data element to which they relate). A Metadata Structure Definition has mechanisms for identifying which metadata are relevant in terms of concepts, rules for its usage (e.g. mandatory/conditional) and code lists.
8. SDMX Technical Specifications prescribe that any kind of metadata report uses concepts to be defined in a “Concept Scheme” which is defined and maintained by a “Maintenance Agency”. A “Metadata Concept Scheme” brings clear advantages both in metadata production and sharing by:
a) improving the way reference metadata are stored; b) publishing a non-redundant and extended list of metadata concepts used; c) setting the requirements for a concept family of reference metadata to be exchanged with statistical partners and re-used.

III. THE SDMX CONTENT-ORIENTED GUIDELINES

9. The SDMX initiative recently delivered a set of draft "content-oriented guidelines" which have been posted on the web (at <http://www.sdmx.org>) for public comment until 31 May 2006. These guidelines contain recommendations for creating interoperable data and metadata sets using SDMX standards. The work, focused on the harmonisation of a limited range of 26 high-level concepts common to a large number of statistical domains, is aimed at encouraging the exchange of comparable statistical information both between international organisations and between national agencies and international organisations.
10. The SDMX content package emphasises the identification of reference³ metadata concepts and subject-matter domains to test SDMX standards. The package also includes a newly revised

³ In SDMX, "reference" metadata are metadata describing the contents and the quality of the statistical data, normally including "conceptual" metadata, describing the concepts used and their practical implementation; "methodological"

version of the Metadata Common Vocabulary (<http://www.sdmx.org/projects/project.aspx?id=4>) which includes definitions consistent with existing international standards and guidelines, with the terminology used within SDMX and, to the extent possible, in other related projects. The MCV, which flows from initial Eurostat-OECD efforts on the joint development of statistical glossaries, is focused on a system of definitions for metadata concepts which can be used for any statistical domain and independently from any general model. The Vocabulary is only concerned with the elaboration of building blocks, subject to terminology standards, easily understandable and re-usable: agreement on a basic vocabulary of concepts still provides each agency with the flexibility of deriving a variety of specific formats and models according to its specific needs. The list of terms and associated definitions simply provides a common language applicable across domains.

11. The MCV plays an important role in providing the common set of terms and definitions that can be used to describe the data. But the standardisation also includes, where appropriate, the representation of concepts with code lists and the identification of the role they play within data and metadata structures. Further work will be undertaken at the end of the current public consultation process to produce a consolidated version of the MCV to ensure that all definitions in the Vocabulary are in line with the agreed list of cross-domain concepts, to improve the content and wording of MCV definitions, and to make available the SDMX-ML version of the glossary.
12. Eurostat and OECD currently use the MCV to ensure clarity and terminological consistency within both organisations' respective metadata repositories (Eurostat free dissemination and OECD MetaStore)⁴. The use of standard definitions taken from the MCV is similarly encouraged within metadata-related projects and activities of other international organisations and national agencies. The availability of a web repository of standard metadata definitions, available for all Internet users, is also a unique chance for creating a common understanding across countries, for instance across European Union or OECD Member countries.

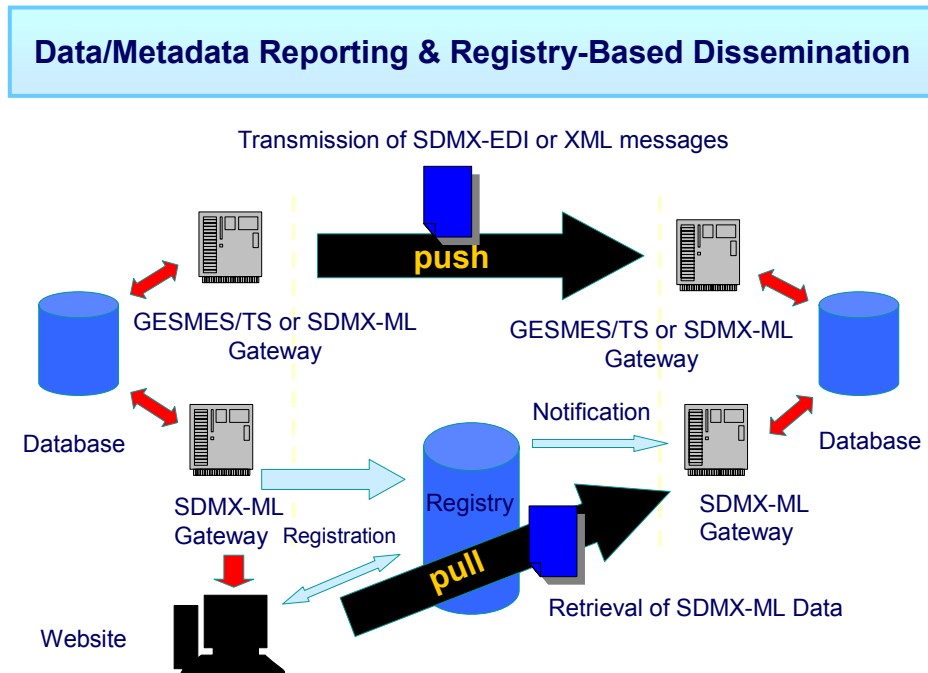
IV. FACILITATING THE EXCHANGE OF METADATA THROUGH SDMX: HOW NATIONAL AGENCIES CAN BENEFIT FROM THIS

13. Through the alignment to SDMX standards, there is a concrete possibility of setting the requirements for a concept family of metadata to be exchanged and shared among countries and international organisations. Alignment does not necessarily entail the direct adoption of precisely the same concept by each agency. Although such adoption would facilitate the ability to exchange metadata (with the same content) between agencies, it would be sufficient for organisations to be able to map the granular concepts developed to meet their own needs to a list of high-level cross-domain concepts, as envisaged in the mapping between Eurostat and SDMX concepts presented in annex. Agreement on a common set of concepts by national agencies and international organisations would represent a significant step forward: in this context, Eurostat (and OECD) could play a role in interconnecting European and national metadata for a wide range of indicators of common interest.
14. The technical mechanism for the actual exchange of metadata between organisations is beyond the scope of the current paper. However, the adoption of a common set of concepts would also facilitate direct access by international organisations ("pull" mechanism) in lieu of the physical transmission by national agencies of different metadata to different international organisations ("push" mechanism) as described in the following chart. Incidentally, a data sharing model demands a good coordination between national and international organisations, to make sure that standardised metadata covering a range of common requirements are made available through the

metadata, describing methods used for the generation of the data (e.g. sampling, collection methods, editing processes); and "quality" metadata, describing the different quality dimensions of the resulting statistics (e.g. timeliness, accuracy). These metadata are often stored in a separate metadata repository and they are referenced from the related data element.

⁴ See M. Pellegrino - D. Ward, "Using the MCV terminology for mapping metadata from different institutions: the case of Eurostat and OECD", WP 13, Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS), Geneva, April 2006, available at <http://www.unece.org/stats/documents/ece/ces/ge.40/2006/wp.13.e.pdf>.

web, while additional information is created directly at supranational level, when this is needed to document the data sets which are disseminated.



15. A number of SDMX implementation projects is currently in place, such as the Joint External Debt Hub, the OECD/UNSD Joint Trade Project, the National Accounts World-Wide Exchange (NAWWE), the ECB-Eurosystem joint dissemination of statistics and SODI (SDMX Open Data Interchange). All these projects demonstrate the feasibility of SDMX standards and tools and, at the same time, the fact that the main issues to tackle are statistical, rather than technical: agreement on the use of common and standardised “Data Structure Definitions” and “Metadata Structure Definitions” is strategic for coordinating data and metadata exchange and for reducing in the end the reporting burden on countries.
16. While working on the technical infrastructure, Eurostat is therefore trying to improve the granularity of the metadata format used, with the aim of extending the conceptual coverage and incorporating more elements on quality assessment, according to the criteria identified by the European statistical code of practice. The draft list of granular concepts (described in annex) is built on the current format used, with some limited extensions on quality elements which need to be further detailed. The current list is going to be used for testing the possibility of disseminating a good selection of reference metadata, together with European member States and in coordination with statistical partners, with regard to the SODI project, for the Principal European economic Indicators.

EUROSTAT – SDMX CROSS-DOMAIN CONCEPTS MAPPING

EUROSTAT DISSEMINATION METADATA CONCEPTS		MAPPING TO CURRENT DRAFT OF SDMX CROSS-DOMAIN CONCEPTS
Top level	Child level	
Metadata Update	Last certified without update	Date of update
	Last update of content	Date of update
Contact	Organisation	Contact
	Address	Contact
	Contact name or service	Contact
	e-mail address	Contact
Data coverage	Short description of data domain	Data presentation
	Data breakdown and main variables	Data presentation
	Units of measure	Data presentation
Periodicity	Periodicity of compilation	Frequency and periodicity
	Database frequency	Frequency and periodicity
Timeliness and punctuality	Timeliness	Timeliness and punctuality
	Punctuality	Timeliness and punctuality
Transparency of practices	Legal acts, reporting requirements	Institutional framework
	Rules on confidentiality	Institutional framework
	Internal access	Transparency
	Commentary on the occasion of release	Transparency
Accessibility	Notification of changes in methodology	Transparency
	Release calendar	Release calendar
	Simultaneous release	Simultaneous release
Quality cross-checks	Dissemination formats	Dissemination formats
	Documentation on methodology	Accessibility of documentation
	Related data and quality cross-checks	[No direct concordance]
	References to quality reports	[No direct concordance]
Accuracy and reliability	Overall accuracy assessment	Accuracy
	Quality checks before release	Accuracy
Comparability and coherence	Comparability over time	Comparability and coherence
	Comparability over space	Comparability and coherence
	Comparability with related sources	Comparability and coherence
	Comparability between datasets	Comparability and coherence
Relevance	Breaks in time series	Comparability and coherence
	Rate of available statistics (user needs)	Relevance
	Intended audience and purpose	Relevance
	Supplementary data	Supplementary data
Statistical concepts and classifications	Statistical concept	Statistical concept
	Definition of indicators	Statistical concept
	Classification system	Classification systems
	Conformity with official standards	Classification systems
	Classification coverage	Classification systems
Scope of the data	Reference area / geopolitical entity	Scope/coverage
	Time coverage	Scope/coverage
	Statistical unit	Scope/coverage
	Statistical population	Scope/coverage
Accounting conventions	Reference period	Accounting conventions
	Base period	Accounting conventions
	Basis for recording	Accounting conventions
Nature of basic data	Data source used	Source data
	Type of survey	Source data
	Methods of data collection	Source data
Compilation practices	Compilation	Statistical processing
	Adjustments and weights	Statistical processing
	Data validation	Statistical processing
	Revision policy and practice	Revision policy and practice
Other	Warnings on re-use and limitations	[No direct concordance]