SDMX implementation in the statistical practice

Marco Pellegrino, Eurostat [1]

Summary

Keywords: IT, metadata, registry, SDMX

The SDMX initiative (Statistical Data and Metadata eXchange) is aimed at developing efficient processes for the exchange and sharing of statistical data and metadata among international organisations and member countries. For this purpose, SDMX produced a set of technical standards and statistical (content-oriented) guidelines. SDMX technical standards are recognised as ISO Technical Specifications 17369 in the version 1, while version 2 is in the process of ISO validation, and is recognised as the preferred standard for data exchange at UN level. In January 2009, SDMX has also disclosed the first package of content-oriented guidelines, recommending practices for creating interoperable data and metadata sets using SDMX technical standards.
Eurostat is currently implementing the SDMX content-oriented guidelines for the harmonisation of both structural metadata (standard data structures and standard code lists) and reference metadata (standard structure for explanatory texts) in several domains.

This paper also reports on the development of some SDMX-compliant tools and on the general IT infrastructure. SDMX envisages the promotion of a data-sharing architecture using the pull mode to facilitate a low-cost and high-quality exchange of data and metadata: a data reporting organisation publishes data once, and lets their counterparties "pull" data and related metadata as required.
Based on past experiences, Eurostat launched new actions to improve the development of SDMX within the Member States, by providing re-usable software components aimed at implementing a complete SDMX architecture for data providers; this also includes technical advice in order to design correctly the required systems. This action on IT components, which is the object of the SDMX Technical workshop scheduled in Madrid on 22-23 September 2009, can be considered as a sort of "reference guide" for the ESSnet multi-beneficiary grant on SDMX which is just about to start for the initial period of 12 months.

1. Background

The SDMX initiative is aimed at developing efficient processes for statistical data and metadata exchange among international organisations and their member countries. In the years 2001-2008, SDMX produced a set of technical standards and statistical (content-oriented) guidelines. SDMX technical standards are recognised as ISO TS (technical specifications) 17369 and are recognised as the preferred standard for data exchange at UN level. Full information on the SDMX standards, guidelines and organisation is available at http://www.sdmx.org.

The version 2.0 of SDMX introduced, in 2005, a series of enhancements on the previous version, in particular on metadata management (with the introduction of the “metadata structure definition” to describe the structure of a metadata set) and with the so-called registry architecture, useful for providing visibility to large amounts of data and metadata. The SDMX Registry can be seen as a central application which is accessible to other programs over the Internet to provide information needed to facilitate the reporting, collection and dissemination of statistics.

[1] Eurostat, Unit B5, Statistical Information Technologies (marco.pellegrino [at] ec.europa.eu).

2. SDMX content-oriented guidelines

Content-oriented guidelines 2009: a step forward towards statistical standardisation

In January 2009, SDMX sponsors disclosed the first package of content-oriented guidelines (COG). This package recommends practices for creating interoperable data and metadata sets using SDMX standards. They focus on the standardisation of specific concepts and terminology that are common to a large number of statistical domains. Such harmonisation is useful for achieving a more efficient exchange of comparable information, building on the experience already gained in several implementations, so that IT systems which exchange data and metadata may understand what the data or metadata refer to, without any big problem in determining the semantic equivalence between concepts. Compared to previous drafts of the SDMX COG, the 2009 release is broader in scope and also reached a better quality of contents.

The SDMX COG package, available on the SDMX website, comprises the following main elements:

- Cross-Domain Concepts: a list of metadata concepts relevant to several statistical domains and recommended for use in data and metadata structures to promote standardisation and interoperability. The composition of this list was guided by concepts already in use by SDMX sponsoring organisations or other statistical organisations.
- Cross-Domain Code Lists: an initial set of 9 code lists recommended for use, describing cross-domain concepts such as observation status, confidentiality, frequency, sex, reference area and currency. These lists contain code values and descriptions which are recommended by SDMX or against which the codes in use by statistical organisations can be mapped (although statistical organisations may dispose of additional code values and descriptions which are not included in the lists presented).
- Statistical Subject-Matter Domains: a standard structure of statistical domains along three main categories: demographic and social statistics, economic statistics and environment, and multi-domain statistics. This subject-matter domain structure provides orientation for organising the production and exchange of statistical data and metadata within SDMX.
- Metadata Common Vocabulary: a repository containing 397 concepts and related definitions which should be used as main basis when dealing with metadata. This vocabulary also covers the cross-domain concepts list.

The implementation of the SDMX content-oriented guidelines for metadata

Eurostat is currently implementing the SDMX content-oriented guidelines across Eurostat for the harmonisation of both structural metadata (standard data structures and standard code lists) and reference metadata (standard structure for explanatory texts).
With regard to reference metadata, Eurostat is using the cross-domain concepts list for the implementation of a Europe-wide standard (ESMS, Euro-SDMX Metadata Structure) to be used for metadata reporting to Eurostat and for exchanging metadata between international organisations, for instance between Eurostat, the European Central Bank and the IMF.

The ESMS structure, fully compliant with SDMX standards and guidelines and with the European Statistics Code of Practice, aims at documenting methodologies, quality and the statistical production processes in general. It uses 21 high-level concepts, with a breakdown of sub-items, strictly derived from the SDMX list of cross-domain concepts. Most of the reference metadata in the ESMS are currently inserted as free text, although in the near future some of them may even follow a code list (e.g. frequency, or reference area).

Within Eurostat's metadata system, the ESMS has already replaced the SDDS since December 2008. The transition from the old to the new standard has been made easier by the availability of a metadata IT application (EMIS) used for treating, storing and extracting reference metadata, and of course by the conceptual similarity between the SDDS and ESMS templates. On top of the statistical concepts which were already part of the SDDS template, the ESMS contains a number of additional statistical concepts, mainly related to data quality, such as accuracy, comparability, coherence, relevance, etc. Therefore, the ESMS better integrates the information which is part of the ESS standard quality report (including specific quantitative quality indicators, such as non-response rate). The quality information

contained in the ESMS will allow a much better comparison of the quality reached by each statistical survey.

ESMS high-level concepts

1. Contact	**8. Release policy**	**15. Timeliness and punctuality**
2. Metadata update	**9. Frequency of dissemination**	**16. Comparability**
3. Statistical presentation	**10. Dissemination format**	**17. Coherence**
4. Unit of measure	**11. Accessibility of documentation**	**18. Cost and burden**
5. Reference period	**12. Quality management**	**19. Data revision**
6. Institutional mandate	**13. Relevance**	**20. Statistical processing**
7. Confidentiality	**14. Accuracy and reliability**	**21 Comment**

The ESMS, which uses SDMX cross-domain concepts and is supported by a Metadata Structure Definition and by a "generic" SDMX-ML metadata format, is addressed to the whole European Statistical System: Eurostat, Member states and associated countries. The structure has been recommended as a standard format through a formal Recommendation (Commission Recommendation 2009/498/EC of 23 June 2009 on reference metadata for the European Statistical System) proposing that EU Member States and associated countries use the ESMS format when compiling and transmitting domain-specific reference metadata to Eurostat. For more details, see http://intragate.ec.europa.eu/emis/JO-2009-498-EC.pdf. While, in the first step, the ESMS is being implemented within Eurostat, in a second phase also Member States are going to make use of the standard format. This successive implementation should improve the metadata exchange and sharing, reducing redundancies and increasing comparability.

How can national and international agencies take advantage of SDMX for metadata exchange?

Based on the substantial development made on both technical standards and content guidelines, we have now considerably progressed in the implementation of advanced standards for the exchange and sharing of data and metadata, which was the original objective of SDMX.
Nevertheless, with regard to metadata, the objectives of the SDMX initiative can be fully achieved if, together with the set of standard concepts, also shared arrangements supporting the exchange are put in place. The key idea behind an "open metadata interchange" is that, through the use of a common set of statistical concepts, linked to a standard terminology, the set up of a multilateral exchange of reference metadata among countries and international organisations now gets possible. The next working step is that international organisations may agree on a core set of reference metadata concepts, based on the ESMS, on the SDDS/DQAF and other standards. On the basis of such an agreement, the exchange of SDMX-based reference metadata between national and international organisations can be organised in practice.

Using SDMX formats and web services, each participating statistical organisation will be able to identify and retrieve, among the metadata made available, those which are relevant for its own framework. European countries, in this context, could provide metadata to more international organisations once for all, for the same SDMX concept and for the same data set. This does not necessarily require the adoption of exactly the same statistical concepts by each agency for its own metadata system: each organisation just needs to map its own statistical concepts to the list of concepts identified in a common Metadata Structure Definition. This use of SDMX may help to achieve an immediate reduction of the burden at national level, because using common XML formats and standardised web tools, a statistical organisation would also be able to identify and retrieve those metadata which are relevant for its own metadata system. The mapping between the metadata concepts used by Eurostat, IMF and OECD – also present in the COG package – supports the idea of such an international metadata exchange and sharing.

The ESMS is intended to be used for facilitating the direct access to ESS metadata on the web ("pull" mechanism) instead of the current transmission by national agencies ("push" mechanism) thanks to the SDMX registry architecture put in place at Eurostat. The same principle could be used for the transfer of metadata files between Eurostat, the European Central Bank and the International Monetary Fund for the "Euro area" page of the Dissemination Standards Bulletin Board. In this framework, Eurostat and the European Central Bank could coordinate Euro area requirements and metadata flows interconnecting national metadata systems. EU countries, at the end, would provide metadata to more organisations at the same time, for the same metadata concepts, possibly using information extracted by their own metadata systems, reducing manual interventions, double work and inconsistencies.

3. IT infrastructure and tools for supporting the SDMX implementation

From a general point of view, SDMX promotes a data-sharing architecture using the *pull* mode to facilitate low-cost and a high-quality statistical data and metadata exchange: a data reporting organisation publishes data once, and lets their counterparties "pull" data (and related metadata) as required. The data-sharing architecture is based on the possibility of discovering easily where data and metadata are available and how to access them. In general, SDMX tools are produced in a decentralised manner and made freely available to the user community on the web (a specific section dedicated to IT tools has been added to the SDMX website at http://sdmx.org/?page_id=13).

In the framework of the X-DIS project (XML for Data Interoperability in Statistics) financed by the Commission programme IDABC, Eurostat has developed a number of IT applications supporting the implementation of SDMX. These are available for Member States as Open Source Software under the EUPL Open Source Software licence, as packages containing the application, documentation and source code files; the download links are available from CIRCA and will also be added to the Open Source Observatory and Repository, OSOR (www.osor.eu). Other tools have been developed for internal use at Eurostat: these are not portable, because more or less closely tied to Eurostat IT infrastructure.

Portable software packages

The following tools were developed as portable packages or components which can be installed in other organisations:

- SDMX Registry – a metadata registry which implements the SDMX specifications. This application provides a web-based user interface as well as web services for interacting with the structural metadata objects in use within Eurostat and with statistical partners. It will enable NSIs and other external organisations to obtain metadata such as Data Structure Definitions (DSD), Metadata Structure Definitions (MSD) and Code Lists. The Eurostat SDMX Registry currently installed is populated with the all the SDMX artefacts currently available within Eurostat.
- Data Structure Wizard (DSW) – a desktop application designed to work with SDMX-compliant registries for editing and viewing structural metadata. The Data Structure Wizard can be used both off-line and on-line, depending on user choices and access rights. The off-line mode is intended to be used for the creation and maintenance of SDMX objects. In the on-line mode, users can interact with any standard-compliant SDMX registry.
- SDMX Converter – application offering the ability to convert between all the existing formats of the SDMX version 2.0 standard (generic, compact, utility and cross-sectional) as well as GESMES/TS, GESMES/2.1, GESMES/DSIS (SDMX-EDI 2.0) and CSV formats. It can be used as a service in a “service-oriented architecture”, called from a command line script, or the relevant Java classes could be linked into other programs.
- Business Cycle Clock version 2.0 – an interactive application for dynamic visualisation of short-term economic indicators, which can be fed by data in SDMX-ML format. It is now integrated into Eurostat's web site and a community version is available as OSS in the OSOR repository.
- X-DIS Visualization Tool version 1.0 – transforming SDMX-ML data (generic and compact SDMX-ML messages) into readable tables in HTML format, using XSLT (Extensible Stylesheet Language Transformation).

Tools for reference metadata

Other tools are being developed to facilitate the delivery by Member States of SDMX reference metadata. These allow Member States to prepare reference metadata files using an online editor linked to the SODI infrastructure at Eurostat, or offline using a standard template. The reference metadata tools will soon be available but are still at the prototype and test stage:

- The National Reference Metadata Editor deals with national metadata: through this tool, national producers who are not in a technologically advanced state of producing SDMX-ML files directly, can produce domain-specific metadata and send them to Eurostat via the "single entry point" in SDMX format. The Reference Metadata Editor uses a standard web questionnaire based on metadata concepts and report structures, such as the Euro-SDMX Metadata Structure. It is installed as part of the Eurostat IT environment and will be accessible to Member States as a web application.
- The SDMX Metadata Template, built in the Metadata Editor, will be also usable in off-line mode for the compilation of national reference metadata.

How can Member States use the Eurostat SDMX Registry and the related tools

The SDMX Registry application performs a number of tasks:

- Providing information about the structure of datasets and reference metadata sets, answering questions like: what code lists do they use? What concepts are involved?
- Providing information about what datasets and reference metadata sets are available, and where they are located;
- Providing information about how the datasets and reference metadata sets are provided: how often they are updated, what their contents are, how they can be accessed, and similar questions;
- Allowing applications to sign up (or subscribe) for notifications, so that when information in the registry is updated (for example, a dataset or reference metadata set of interest becomes available, or structural metadata are changed) the application is automatically alerted.

Since August 2008, the first version of the registry was installed and running in the European Commission Data Centre. The registry can be used to obtain DSDs and MSDs, to obtain other harmonised structural metadata, such as code lists, and to visualize those structural metadata. The graphical user interface is accessible at https://webgate.ec.europa.eu/sdmxregistry. In order to login, a CIRCA user-id and password and the domain "circa" (in lower case) must be used. The registry already contains some of the most important structural metadata, but Eurostat is currently adding further content, with the aim of including all Eurostat and ECB DSDs, code lists and MSDs.

The Eurostat SDMX Registry provides the registry web services according to the SDMX Version 2.0 standards. The web services are planned to be accessible to applications outside the Commission starting from 2010. Eurostat has also deployed a training instance of the Registry, which can be used as a "sandbox" for training courses and presentations, without the risk of modifying the real registry. Access to the training environment can be provided on request.

The Eurostat SDMX Registry is intended to be accessible both for human users (via the user interface) and for applications. If needed, NSIs could use the portable version of the SDMX Registry application as a basis for setting up their own in-house registries. In addition, users can download the Data Structure Wizard application and use it in off-line and on-line mode.

The Data Structure Wizard application could be used by Member States in the following ways:

- In on-line mode, as a tool to view DSDs and other structural metadata held in the Eurostat SDMX Registry. It should be noted that, for certain purposes, the Data Structure Wizard may offer a more user-friendly interface than the registry;
- In off-line mode, as a tool to support learning about SDMX, for example by allowing users to experiment with the creation of DSDs. Of course, locally-created DSDs will not be loaded into the Eurostat SDMX Registry.

Re-usable software components for implementing SDMX in member States

SDMX standards and IT architecture, together with content guidelines, are being used for a series of implementation projects at Eurostat, also involving ESS countries. One of these is the European Census Hub for the dissemination of the result of the 2011 census data in the European Union, via

Eurostat web site. Others are the SODI project, the Demography Rapid Questionnaire and the Euro-Groups Register (EGR). These experiences have highlighted some important points:

- Building the SDMX architecture, from a data producer point of view, requires the analysis of several factors, and consequently the development of complex software modules.
- National data are often stored in Member States' repositories, and described differently from how they could be in the SDMX Data Structure Definitions defined at international level;
- The start-up phase is crucial, because the expert knowledge of SDMX standards, XML and related technologies (e.g. web services) is not easily available;
- The exchange of know-how and software between national institutes – encouraged by Eurostat (see Census Hub) – has allowed in some cases a quicker development of the IT infrastructure.

Based on those points, a new project has been launched to improve the development of SDMX within the Member States, by providing technical advice and re-usable software to those which are interested in implementing a complete SDMX architecture for data providers. The action also includes technical advice in order to design correctly the required systems.
This action will also include the analysis of selected NSIs' architectures, with a particular attention for solutions which are already shared in the statistical community (e.g. pc-axis) and the inventory of existing software developed at national level for SDMX projects, together with proposals on how to integrate them in the SDMX reference architecture. This is also intended to facilitate and coordinate multilateral IT developments among Member States, exploiting the architecture and components already developed by Member States which participate in the SODI and Census Hub projects.

Detailed specifications for the SDMX reference architecture, together with the definition of single components and the interfaces between building blocks will be made available on CIRCA. Software components will be made available as Open Source Software as packages containing the application, documentation and source code files. The download links are available from CIRCA and will also be added to the Open Source Observatory and Repository.

This short-term action is presented at the SDMX Technical workshop scheduled in Madrid on 22-23 September 2009 (documentation available from CIRCA in the "SDMX development" directory at http://circa.europa.eu/Public/irc/dsis/sdmxdevelopment/library?l=/&vm=detailed&sb=Title) and can be considered as a sort of "reference guide" for further developments. In this context, the project - in addition to producing short-term results for on-going projects such as the Census Hub – is also going to provide useful additional material that will facilitate the work of the SDMX ESSnet multi-beneficiary grant which is just about to start for an initial period of 12 months.

Training and other capacity-building actions

Training has been a central part of the implementation strategy for SDMX in the last years. In 2007, 2008 and 2009, several training courses on SDMX were held or are planned, either for statisticians or for system developers, for Eurostat staff members or for external people. These courses are held in addition to the tutorial sessions normally taking place at the meetings of the STNE and Metadata Working Groups.

In line with the priorities of the SDMX Sponsors Committee – and as requested by the UN Statistical Commission – more training and capacity-building actions are being organised. In particular, the following additional activities are planned:

- Preparation of a self-learning package (the first set of e-learning tools will be released in the Autumn through the SDMX webpage)
- Regional workshops coordinated with Eurostat cooperation programmes such as Medstat, Tacis, Cards, etc.
- Specific SDMX workshops within the European Statistical System, such as the SDMX technical workshop in Madrid in September 2009, or specific sessions in several countries.

4. Conclusion

The SDMX technical standards and statistical guidelines have reached a good level of maturity and are now ready for use and implementation by statistical organisations. The reference architecture and IT tools developed by Eurostat provide a good basis for the use of SDMX, and in this sense they are

currently used for supporting an implementation plan within Eurostat. Still, this is not enough for implementing the IT infrastructure in many member States. From a data producer point of view, building the full SDMX architecture requires the analysis of several factors, the development of software modules, and a mapping between concepts stored in existing databases and those described in the Data Structure Definitions to be used for data exchange.
The possibility of supporting countries in designing their IT infrastructure – through dedicated implementations and with participated actions such as the ESSnet grants – is, therefore, the new challenge for the implementation of SDMX at national level in 2010-2011.

5. References

- SDMX Content-Oriented Guidelines, January 2009 (http://www.sdmx.org/index.php?page_id=11)
- Götzfried, Pellegrino, "The Implementation of SDMX content-oriented guidelines with regard to metadata exchange within the European Statistical System", SDMX Global Conference, Paris, January 2009 (http://www.oecd.org/dataoecd/41/8/41648971.pdf?contentId=41648972)
- Lindblad, Rizzo, "Eurostat SDMX Registry", SDMX Global Conference, Paris, January 2009 (http://www.oecd.org/dataoecd/5/19/41682274.pdf?contentId=41682275)
- Eurostat – Unit B5, "SDMX Reference Architecture", STNE working group, June 2009, http://circa.europa.eu/Public/irc/dsis/stne/library?l=/meeting_stne_16-17/sdmx-2009-33_architectur/_EN_1.0_&a=i