European Conference on Quality in Official Statistics (Q2014)
Vienna, 3-5 June 2014

Better data quality through global data and metadata sharing

Agne Bikauskaite, Luca Gramaglia, August Götzfried, and Håkan Linden

Eurostat, European Commission, L-2920 Luxembourg

Keywords: SDMX, DSD, registries, business processes, harmonisation, and metadata

1. Introduction

International data and metadata sharing means the exchange of data and/or metadata in a situation involving the use of open, freely available data formats and where process patterns are known and standard. In data sharing exchange, any organization or individual can use any counterparty's data and metadata (assuming they are permitted access to it). This model requires that data and metadata providers and consumers adhere to the standards.[1]

Currently, international organisations are embarking more and more international data and metadata sharing arrangements, mainly for reasons of cost savings and of increase of quality of statistical output.

This paper describes Eurostat's data and metadata sharing strategy and how it contributes to the quality of European statistics and to the improved exchange and use of data and metadata in the international context beyond the European Statistical System (ESS).

2. The context of global data and metadata sharing

Data and metadata sharing is the set of actions and policies by which data and the metadata necessary for their interpretation are exchanged between statistical organisations.

The basic objectives for global data and metadata sharing are:

- To reduce the national reporting burden to international organisations;
- To use the resources of national and international organisations more efficiently;
- To ensure that the respective data and metadata disseminated by international organisations are identical and respond to the quality requirements set (including e.g. the frequency and timeliness needed);
- To improve the dissemination of internationally harmonised and consistent data and metadata.

[1] See also the SDMX Metadata Common Vocabulary which is part of the SDMX Content-oriented Guidelines, www.sdmx.org

In a number of statistical domains (e.g. National Accounts, Balance of Payments statistics, etc.), international organisations start to move from multiple data exchange flows to more integrated and streamlined data exchange between national and international organisations. The necessary pre-condition is to use internationally agreed statistical and technical standards as well as a sophisticated IT infrastructure for these redesigned processes. SDMX (Statistical Data and Metadata Exchange) provides these standards and this infrastructure. The redesigned exchange and dissemination processes lead to quality improvements of the data and metadata sharing.

3. The Standards, IT infrastructure and tools

For this purpose, Data Structure Definitions[2] (based on SDMX) have to be agreed upon between national and international organisations. These globally agreed Data Structure Definitions (DSDs) have to contain harmonised statistical concepts and harmonised structural metadata.[3]

The SDMX based DSDs and Metadata Structure Definitions[4] (MSDs) contain dimensions and attributes. These dimensions and attributes contain the harmonised statistical concepts and the harmonised code lists.

3.1 Harmonised structural metadata

The code lists which are common for many statistical domains are defined as Cross-Domain. As example, could be taken a code list dealing with age classes. To express concepts like "From 15 years to 20 years excluding 16 years", "Over 30 years", etc. the following set of standard operators is proposed:

- "T" for expressing ranges;
- "_" for the combination of two codes;
- "X" for expressing "except" or "excluding";
- "GT" for "greater than", "LE" for "equal to or less than", etc.

[2] A Data Structure Definition is defined as follows: Set of structural metadata associated to a data set, which includes information about how concepts are associated with the measures, dimensions, and attributes of a data cube, along with information about the representation of data and related descriptive metadata; see SDMX Metadata Common Vocabulary, www.sdmx.org.

[3] Structural metadata are those metadata acting as identifiers and descriptors of the data, such as names of variables or dimensions of statistical cubes. Data must be associated to some structural metadata, otherwise it becomes impossible to properly identify, retrieve and browse the data.

[4] A Metadata Structure Definition (MSD) identifies what metadata concepts are being reported, how these concepts relate to each other (typically as hierarchies), what their presentational structure is, i.e. how they may be represented (as free text, as coded values, etc.), and to which formal object they are attached.

In this case the code of the label “From 15 years to 20 years excluding 16 ears” would be Y15T20X16.[5]

These SDMX statistical standards are growing with the increasing number of SDMX DSDs. In any case, when new needs for data exchange at national or international level appear then the existing code lists and concepts should be reused to the largest extent possible. Additions to existing code lists or concepts might get necessary if new data structure definitions are designed for additional statistical domains.

The full list of SDMX Cross-Domain Concepts and related code lists are available on the SDMX website (http://www.sdmx.org). In addition examples are provided of how these concepts and code lists can be used.

Besides Cross-Domain concepts and code lists, Shared or Specific artefacts[6] are frequently used and needed. Specific means that concepts or code lists are used in one statistical domain only, while Shared artefacts are used in a limited number of statistical domain (e.g. in two or three related statistical domains such as National Accounts (NA) and Balance of Payment (BOP) statistics).

3.2 The SDMX Global Registry

The SDMX Global Registry is the central IT application that contains the SDMX DSDs and other SDMX objects (such as code lists) supporting the world-wide implementation of those data structure definitions. It is owned and maintained by the SDMX Sponsors.[7]

The SDMX Global Registry will contain more and more SDMX global DSDs and their related objects as they become available. It is the central reference point and authoritative source for SDMX global DSDs and related objects.

In order to facilitate the implementation of the SDMX standards, the SDMX Global Registry will provide an easy access to the SDMX DSDs and other SDMX artefacts that are used and maintained by national and international organizations and that are used for global data sharing. For more details please see https://registry.sdmx.org/about.html.

[5] For more details on the recommendations for the creation of the code lists see the “Guidelines for the creation and management of SDMX cross-domain code lists”, www.sdmx.org

[6] Artefact is a basic element in the SDMX model from which other elements are derived. Artefacts provide features which are reusable by derived elements to support horizontal functionality such as identity, versioning etc.

[7] The SDMX Sponsors are: BIS, ECB, Eurostat, IMF, OECD, UN, the World Bank.

3.3 The IT infrastructure

In addition to the SDMX based DSDs and the SDMX Global Registry also an SDMX IT infrastructure is needed for international data sharing. This concerns in particular the SDMX Hub IT infrastructure (including SDMX Reference Infrastructure (SDMX-RI) and the SDMX web services) which should ensure the timely and efficient international data exchange[8].

4. Implementing Eurostat's data sharing strategy

4.1 Implementing the data sharing in statistical domains

The data and metadata exchange process of several statistical domains in Eurostat is being restructured to comply with Eurostat's data sharing strategy. Some of the implementation projects involve also other international statistical organisations, while others are internal to the ESS.

At international level, a flagship project for the implementation of Eurostat's SDMX-based data sharing strategy is the implementation of SDMX for National Accounts, Balance of Payments and Foreign Direct Investment statistics. Conducted in collaboration with the other SDMX Sponsors organisations with the aim of streamlining the international exchange of essential macro-economic data, the project has led to the creation of seven internationally agreed DSDs to cover the afore-mentioned domains. The first instance of data exchange according to this new paradigm will take place during the third quarter of 2014. In the wake of this experience, Eurostat is participating in international SDMX implementation in other statistical domains, such as Research and Development statistics, Education statistics, and International Merchandise Trade statistics.

At European level, Eurostat has completed SDMX implementation for its data exchange with EU Member States in five statistical domains. Another ten domains are undergoing SDMX implementation for data exchange. As far as metadata is concerned, SDMX is being or has been implemented to normalize metadata exchange between EU Member States and Eurostat in over 20 statistical domains.[9]

[8] See SDMX IT tools on www.sdmx.org and the Eurostat SDMX Info Space where a series of SDMX IT tools are made available (e.g. the SDMX Converter, the Mapping assistant etc.); see also: https://webgate.ec.europa.eu/fpfis/mwikis/sdmx/index.php/Main_Page

[9] The complete list of domains is available at https://webgate.ec.europa.eu/fpfis/mwikis/sdmx/index.php/Main_Page

4.2 Development of the technical infrastructure

In order to successfully implement its SDMX-based data sharing strategy, Eurostat has invested heavily in the development of the relevant technical infrastructure. While efforts have been devoted to the creation of software to facilitate the creation of SDMX artefacts and to ease the conversion between different data formats, work has concentrated on the development of two key components for a successful SDMX strategy: the SDMX registries and the SDMX-RI.

As was mentioned in section 3.3, an SDMX Registry provides a facility in which SDMX artefacts can be stored and queried. SDMX Registries play a capital role in Eurostat's vision for data sharing, as they allow for the structural metadata underlying data and metadata exchange to be centrally managed and to be made available via Web services: production systems and applications can therefore access the relevant metadata and use it to drive statistical data processing. Eurostat has been involved in the creation of two separate SDMX Registries: the Euro-SDMX Registry, which serves as the central repository for all SDMX artefacts used solely for ESS-internal data exchange, and the Global SDMX Registry, developed in collaboration with the other SDMX Sponsor organisations, which plays the same role for the SDMX objects created in the framework of international or global implementation projects (see section 3.2).

The SDMX-RI is a generalized service infrastructure which allows the exposure of data in existing databases to SDMX Web Services. Use of the SDMX-RI opens the possibility to use new approaches to data exchange by allowing for data to be "pulled" by the data receiver (i.e. queried from the data provider's database) rather than "pushed" to the data receiver (as has traditionally been the case). The SDMX-RI has been successfully installed in the vast majority of EU Member States in the framework of the Census-HUB project. Many Member States are also considering its re-use in other statistical domains undergoing SDMX implementation, most notably National Accounts.

4.3 Development of a project management methodology

Eurostat's experience in applying its data and metadata sharing strategy has highlighted that SDMX implementation projects carry a high degree of complexity. The complexity stems from two intrinsic characteristics of these projects:

- Any SDMX implementation project encompasses both conceptual (statistical) and technical (IT) aspects. This implies active cooperation and communication between subject-matter and IT experts.
- As SDMX is a standard for data and metadata *exchange*, SDMX implementation always relies on close collaboration between all the partners involved in the exchange. This entails need to take into account the needs of different stakeholders and to work across borders and organisations.

It must be noted that the sources of complexity cited above are not linked to the intricacy of the SDMX standard itself. They are rather inherent to the very nature of international data sharing and would in no way be solved by substituting SDMX with another standard.

In order to cut through the complexity, Eurostat developed over the years, in collaboration with other international organisations involved in SDMX, a solid project management methodology to underpin every SDMX implementation project. The project management methodology consists of two main elements:

- A structured implementation process: based on the experience gained in recent years, Eurostat has established a four-phase process for SDMX implementation.[10] Strict adherence to the developed approach ensures a correct and efficient tackling of the issues encountered during the course of SDMX implementation.
- The SDMX project manager: in order to ensure a good balance between subject-matter and IT concerns in every SDMX implementation project, Eurostat has established the figure of the SDMX project manager. The SDMX project manager is neither an expert in a specific statistical domain nor an IT developer: his role is to model the data exchange process according to the SDMX standards and to guide the collaboration between and within different statistical organisations.

The application of this project management methodology has resulted in a significant speed-up of SDMX implementation both at international and European level in the past few years compared to the first period of SDMX implementation in Eurostat (2009-2010).

5. The simplification of international data and metadata flows

The national and the international organisations involved in statistical data and metadata sharing have to collaborate. For the high quality data exchange common terminology, rules

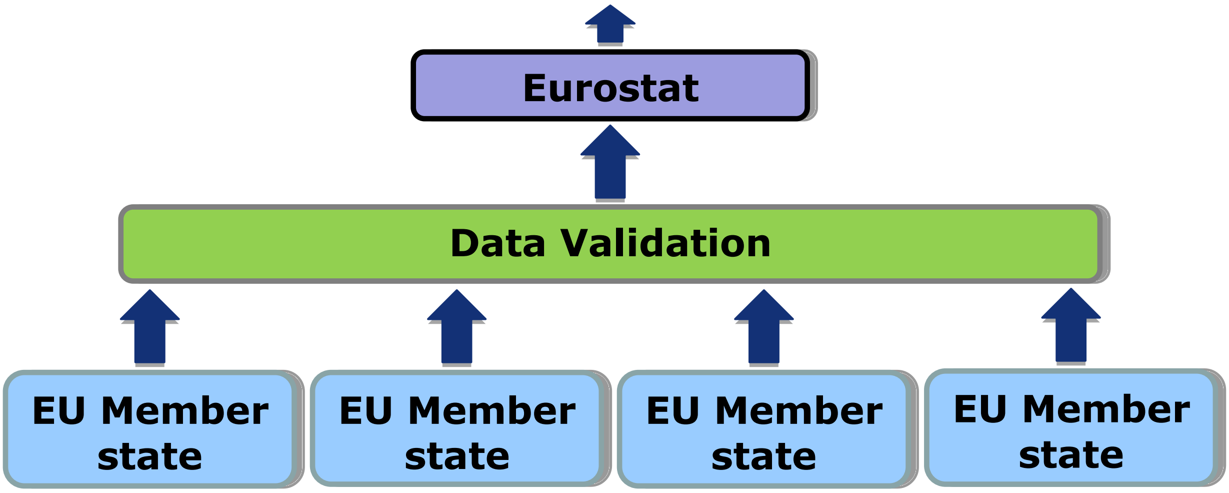
[10] A more detailed description of the implementation process model is available at https://webgate.ec.europa.eu/fpfis/mwikis/sdmx/index.php/Main_Page

and procedures have to be applied. One leading international organisation acts as global hub of data exchange and sharing for a predefined number of countries. To improve the data quality and finally care for identical data dissemination, national data and metadata should only be exchanged with one international organisation and validated once.

5.1 International data sharing: a model

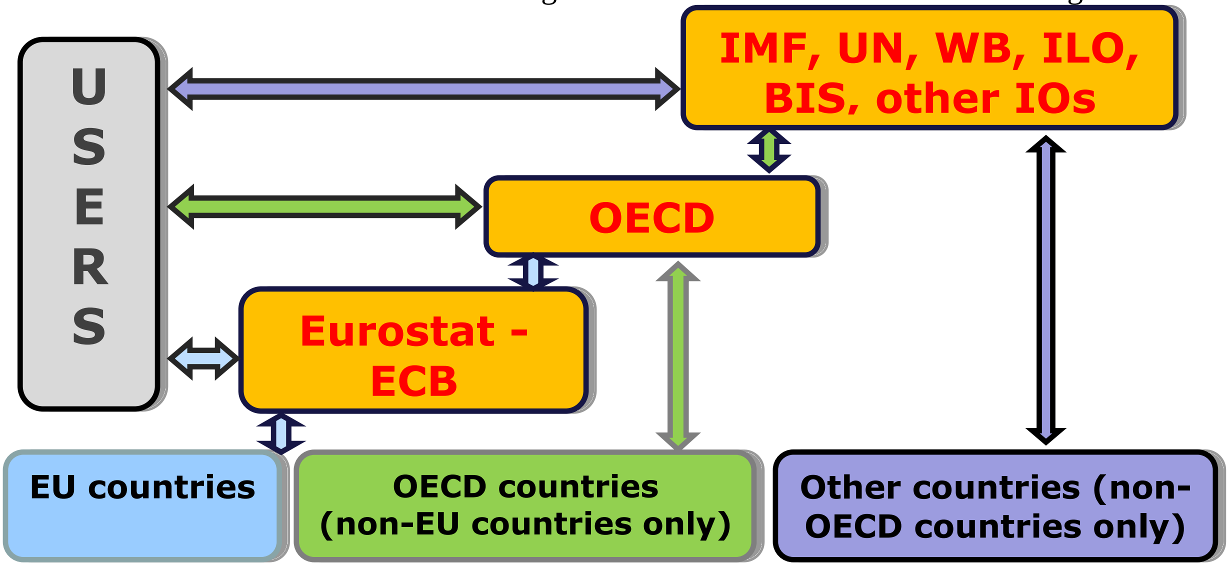
The two following pictures illustrate the exchange of European statistics on the basis of DSDs for global use between EU Members states and Eurostat on the one hand and between Eurostat and other international organisations and users at the other hand.

Picture 1: Data exchange from EU Member states to Eurostat



European Statistics flow: from national to Eurostat

Picture 2: Data exchange from Eurostat to other international organisations


USERS
IMF, UN, WB, ILO, BIS, other IOs
OECD
Eurostat - ECB
EU countries
OECD countries (non-EU countries only)
Other countries (non-OECD countries only)

European as International Hub for European statistics

Overall, one data value should be produced by the national statistical organisation; this data value should be validated once and then be disseminate by all international organisations concerned.

5.2 International reference metadata sharing: towards a model

Beside international data sharing, also international reference metadata sharing is now on the agenda of some international organisations for improving and streamlining also the metadata exchange between national and international organisations. This is becoming more and more important given the needs – for instance – for monitoring financial and economic systems in a consistent way (i.e. to have agreed assessments etc. on the quality of the data and the phenomena measured). This means that as for data the users of official statistics across countries should receive consistent, comparable and timely reference metadata linked to the agreed DSDs for global use.

In the past, international organisations have implemented different standards for in particular their reference metadata on quality (both at national and international level) in order to meet their specific needs.

However, the different quality frameworks were converging in the last years as noted in the UN Guidelines for the Template for a Generic National Quality Assurance Framework.[11] In these Guidelines, the co-existence of different quality frameworks is recognised. The Guidelines also draw to a large extent on the SDMX Metadata Common Vocabulary (MCV) for describing the quality concepts.

For progressing towards international reference metadata sharing, the existing reference metadata frameworks and templates would need to as harmonised as possible. The statistical concepts used in these frameworks and templates (or Metadata Structure Definitions) need then to be defined in a unique manner.

As for statistical data, the SDMX statistical and technical standards as well as the SDMX IT infrastructure (such as the SDMX Registry) provides the basis for progressing towards this harmonisation and integration.

Also recent work within the SDMX Statistical Working Group on revision of the SDMX MCV could already now accommodate the need also for more international reference metadata sharing.

[11] UN National Quality Assurance Frameworks: http://unstats.un.org/unsd/dnss/QualityNQAF/nqaf.aspx

6. Governance agreements

International data sharing also necessitates underlying governance arrangements for maintaining and keeping up-to-date the data or metadata structure definitions or other objects used for exchange and dissemination.[12]

Political agreements are to be concluded between international organisations mandating one international organisation with the full responsibility of data exchange, processing and dissemination of official statistics for a statistical domain and for a set of countries. This means that national statistical organisations are exchanging their data only with this one international organisation which then makes available this data to other international organisations in a pre-defined data structure and with pre-defined delays. The same statistical data value with the identical structural metadata is then disseminated by all international organisations concerned by the respective statistics.[13]

For more and more data sharing actions, international maintenance agreements are therefore concluded defining the owners of the DSDs as well as their maintenance agencies. This assures an internally agreed structured management of changes of the DSDs as well as the planned implementation of changes. These agreements assure that the international statistical business processes are kept up-to-date (based on agreed DSDs) and stay relevant.[14]

7. The New Frontier: a SDMX Validation and Transformation Language

Efficient international data exchange – as described above – does not rely solely on the definition of harmonised structures for data exchange. Also common data validation and processing procedures are required (see also picture 1 above).

However, the description of complex mathematical expressions goes beyond the current scope of the SDMX standards. The development of a standard syntax to express and exchange validation and transformations rules would therefore fulfil an important role in achieving more integration of the statistical business processes.

In order to fill this gap, the SDMX Sponsors took the initiative to define a Validation and Transformation Language (VTL). The aim of VTL is to provide an "active" metadata language for data validation, i.e. a machine-readable language whose instances would be unambiguously interpretable by software components.

[12] See also the "Governance of commonly used SDMX metadata artefacts", www.sdmx.org;

[13] Currently international organisations disseminate different data values due to different data structures, differences in data validation and in data dissemination.

[14] For more details on the governance see also the "Governance of commonly used SDMX metadata artefacts", www.sdmx.org

In order to foster wide adoption, VTL will be made as versatile and generic as possible. This means that it will follow a policy of threefold agnosticism:

- The VTL should be domain-agnostic: The VTL would not be oriented towards the specific needs of a specific statistical domain.
- The VTL should be IT-agnostic: The VTL would not be oriented towards a specific IT implementation, but could be used in several IT applications.
- The VTL should be Information Model-agnostic: The VTL should be founded on an "agnostic" Information Model, which can be unequivocally mapped to SDMX, DDI (Data Documentation Initiative), and GSIM (Generic Statistical Information Model). Therefore, while the VTL initiative is being developed under the lead of the SDMX Technical Working Group, representatives from other standards communities are involved in its creation.

A first version of the VTL is expected by the beginning of 2015. The VTL will build upon previous experiences with meta-languages for validation (such as EXL, developed by Banca d'Italia, and VALS, developed by Eurostat). It should then be made available together with the SDMX statistical and technical standards for business process redesign as described above.

8. Conclusions

Standards are keys in the Eurostat strategy for progressing in terms of statistical business process quality and business process integration. The increasing use of SDMX based statistical standards will improve the quality of the underlying statistical processes as well as of the statistical output. In particular the comparability of data between statistical domains as well as between statistical organisations will be thoroughly increased through the use of globally harmonised metadata.

This should lead in the years to come to much more data and reference metadata sharing, in particular between international organisations. Users will gain much from this sharing and better international data and reference metadata quality should be achieved.

9. References

[1] SDMX web site: http://sdmx.org/
[2] Eurostat's SDMX and Metadata Standards Info Space:
https://webgate.ec.europa.eu/fpfis/mwikis/sdmx/index.php/Main_Page
[3] Implementation of SDMX for National Accounts: http://sdmx.org/?p=1087
[4] Eurostat's SDMX and Metadata Standards Info Space for National Accounts:
https://webgate.ec.europa.eu/fpfis/mwikis/sdmx/index.php/National_Accounts
[5] SDMX Global Registry: https://registry.sdmx.org/home.html