European Conference on Quality in Official Statistics (Q2016)
Madrid, 31 May-3 June 2016

Practical experience in the implementation of data and metadata standards in the ESS

Jan Planovsky[1], Bogdan-Sorin Zdrentu[2]

[1] *Eurostat, European Commission, Luxembourg; Jan.Planovsky@ec.europa.eu*
[2] *Eurostat, European Commission, Luxembourg; Bogdan-Sorin.Zdrentu@ec.europa.eu*

Abstract
The objective of this paper is to share Eurostat's experience in constructing standardised data and metadata structures. Implementing these standards enable seamless data collection, validation and exchange using modern IT tools as well as providing well-structured metadata at the same time. In concrete terms, reference metadata are presented in files based on a standardised format called the Euro SDMX Metadata Structure (ESMS) and quality reports are produced and disseminated in the form of the ESS Standard Quality Report Structure (ESQRS). The paper also introduces the "Single Integrated Metadata Structure" (SIMS), which represents an inventory for all ESS quality concepts, reuniting reference metadata from the ESMS and ESQRS.

Keywords: metadata, standards, experience, SIMS, ESQRS, ESMS, SDMX.

1. Introduction

Eurostat is the Statistical Office of the European Union, based in Luxembourg. It is also a Directorate-General (DG) and as such an integral part of the European Commission. This dual function means that Eurostat plays a unique role in standardisation of data and metadata exchange at the European Statistical system (ESS) and Commission levels. National Statistical Institutes, European Commission services and International Organisations (such as the UN, the OECD, the IMF, etc.) are currently embarking on greater international data sharing, reusing data, mainly to increase the quality of statistical output across organisations but with the added benefit of cost savings [1]. Seamless data exchange and dissemination, machine-to-machine communication and faster accessibility to data via well-defined metadata require, however, the implementation of appropriate data and metadata standards.

Eurostat is directly involved in three different initiatives regarding standardisation:

- At the level of the European Commission, an IT rationalisation exercise was launched in 2011 with the objective of avoiding duplication and of promoting common, reusable and flexible solutions. Eurostat was nominated the domain leader responsible for analysis, databases and statistical data rationalisation. One concrete work stream was focused on the standardisation of data structures, the use of statistical standards and the creation of common repositories.

- As a member of the ESS Eurostat is taking the lead in the implementation of the ESS Vision 2020 [3]. This includes a set of projects and initiatives that are grouped in 5 key areas to address certain challenges for European statistics. Concretely, key area No. 4 entitled "Efficient and robust statistical processes" is dedicated to intensifying collaboration within the ESS by sharing tools, data, services and resources. In this context, the implementation of standards for statistical production is necessary to improve the comparability of statistical outputs. Standards are required to ensure smooth communication in the system and to make process components interoperable.

- Eurostat is also an important actor at the global level. It closely collaborates with the United Nations Statistical Division (UNSD). It sponsors the global SDMX[1] initiative to improve "Statistical Data and Metadata eXchange" and actively participates in the work of High-Level Group for the Modernisation of Official Statistics (HLG-MOS, Modernstats)

[1] For details, see http://www.sdmx.org

which promotes a standards-based modernisation of official statistics under the support of United Nations Economic Commission for Europe (UNECE).

There is a considerable synergy between these three dimensions of Eurostat's activities– the European Commission, the ESS and the UNECE. However, we need to be aware of the potential danger of a duplication of activities and a sub-optimal use of scarce human resources. To try to ensure the avoidance of any parallelism in developing standards and models, Eurostat recently explored the issue in detail [4] when comparing the Vision 2020 and Modernstats projects and initiatives. The resulting paper suggests that further scrutiny is required. The HLG-MOS launched a number of initiatives to develop models, frameworks and tools to support standards-based modernisation, such as the Common Statistical Production Architecture (CSPA) and a new Generic Activity Model for Statistical Organisations (GAMSO). Separately, Annex B of the Vision 2020 [3] introduces a Generic Enterprise Model for Statistics (GEMS). Both GAMSO and GEMS provide a formalised description of statistical enterprises covering supporting processes that are based on the General Statistical Business Process Model (GSBPM). This example suggests that there might be an opportunity to converge to a single model. Other examples were found at a lower, more technical level regarding statistical standards for data exchange and dissemination. For example, whilst the statistical community widely accepts and uses SDMX as a standard for data and metadata exchange, the non-statistical community (e.g. DGs of the European Commission) is barely aware of it and, instead, has frequently been using alternative standards, such as XBRL or RDF[2].

So whilst there is no question of the importance of well-established models and stable standards to guide quality management in statistical production, it is essential that the duplication of models is avoided as well as a proliferation of too many standards. This will

[2] For details, see http://www.xbrl.org and www.w3.org/rdf respectively

only be achieved when proper governance structures are set in place and global statistical actors communicate regularly with each other.

2. Statistical metadata, models and standards

Statistical data without metadata, such as reference area, time period and observation value have no meaning. Other types of metadata such as information on survey methodology, sample size, accuracy, etc. provide an even deeper layer of understanding. In this way, statistical metadata can be seen as necessary for the proper production and interpretation of statistical data. Metadata describe statistical data and - to some extent – processes and tools involved in the production and use of statistical data. Metadata can be divided into two main categories, each performing different, well defined and equally important duties:

- Structural metadata are defined by their close association with data; they help identify and process the data. For example, structural metadata include the titles of the variables and dimensions of statistical datasets, as well as the units of measurement used (e.g. EUR), code lists (e.g. for territorial coding), data formats, etc. A standard code list is a code list that has already been harmonised, such as the list of EU countries with standard 2-letter codes.

- Reference metadata describe statistical concepts, methodologies for the collection and generation of data and information on data quality. They also assist with the interpretation of the data, thereby being strongly content-oriented. Reference metadata are sometimes generated, collected or disseminated separately from the statistics to which they refer. They can also be associated with different levels of data: entire collections, data sets from a given country, or for a data item concerning one country and one year.

The above-mentioned Generic Statistical Business Process Model [2] is used for integrating data and metadata standards and serves as a template for process documentation, for harmonizing statistical computing infrastructures and provides a framework for process quality assessment and improvement.

While models like GSBPM or GAMSO represent guidelines and templates which could be fully or partially followed or even extended, statistical standards represent strict and precise set of rules and formats which should be fully respected to enable their practical use. For example, SDMX is an ISO standard designed to describe statistical data and metadata, normalise their exchange, and improve their efficient sharing across statistical and similar organisations. Implementation of statistical standards is facilitated by concrete IT tools, such as SDMX Global Registry[3]. Common understanding of terminology is supported by glossaries. For example, the SDMX Glossary[4] contains concepts and related definitions (e.g. code lists) used in structural and reference metadata of International Organisations and national data-producing agencies.

3. Eurostat implementation experience

3.1 Quality framework and Single Integrated Metadata Structure

The Single Integrated Metadata Structure (SIMS) is a dynamic inventory of concepts, developed as a follow-up to recommendations from the high-level ESS Task Force on the Sponsorship on Quality (SoQ) concerning quality reporting (Theme III). With the objective of streamlining and rationalizing quality reporting across the ESS and across statistical domains, the SoQ recommended in its Final report in September 2011 that "a single metadata structure should be used to derive both producer-oriented and user-oriented quality reports and should be specified in the Methodological Manual".

As a result, SIMS and its Technical Manual [6] were developed by an ESS Task Force on Quality Reporting, a sub-group of the Working Group on Quality in Statistics. They were endorsed by the ESSC at its meeting of November 2013.

[3] https://registry.sdmx.org/FusionRegistry/overview.html

[4] https://sdmx.org/wp-content/uploads/SDMX_Glossary_Version_1_0_February_2016.docx

With the first experiences from implementing these reporting structures, some small updates have been considered necessary: as examples, new concepts have been integrated and their order sometimes revised, particularly in order to assure the coherence between the European Statistics Code of Practice, SIMS and ESMS.

Concerning the order of concepts, each quality report should now start with the presentation of the data (Statistical presentation with all its sub-concepts) and with the process (Statistical process and its sub-concepts). Thereafter, the Quality management issues should be introduced, followed by the output quality dimensions in accordance with the Code of Practice (Relevance, Accuracy and Reliability, Timeliness and Punctuality, Coherence and Comparability, Accessibility and Clarity). Information on Cost, Burden, Confidentiality and Additional Comments should close the report.

As a central coordinator for knowledge transfer, Eurostat focused its efforts on building comprehensive methodological manuals and established appropriate training. The ESS Handbook for Quality Reporting provides detailed information and practical examples on how to create good quality reports. The Handbook is now under revision and will improve the guidelines on metadata concepts, adding more practical examples and better integrating the concept of general descriptive metadata with quality reports. The revision will also be enhanced with new chapters that tackle the issue of the dissemination of metadata in more modern and dynamic environments (infographics, interactive maps).

Following a strong demand for capacity building, Eurostat is working with two Member States (MSs) to provide training on building better quality reports. The approach being taken is to "train the trainers", so that they can transfer information to the interested members of their organisation.

To support the implementation of standards for metadata and data exchange in the ESS, Eurostat is using grants as an instrument to support MSs with the work they need to carry out to complete and transmit the metadata required. Focus at the national level is on the development of quality reporting, implementing technical and statistical standards and the introduction of solutions to support the work, like the introduction and use of the ESS Metadata Handler for national quality reporting and the SDMX Reference Infrastructure for supporting standardised data exchange.

A first set of grant agreements was signed in 2012. A new call for proposals was launched at the beginning of 2015. New grant agreements were signed in September 2015 and the majority of these will run for a period of two years.

The outcome of the first set of grants was the adoption of the SIMS standard for reference metadata and the development of national reference metadata systems across several members of the ESS.

3.2. Metadata reporting structure and IT tool

In order to facilitate the adoption of the metadata standards within the ESS, Eurostat developed an IT platform – the ESS Metadata Handler (MH) - which is used by MSs and Eurostat to create shared metadata files. Building on the success of this platform, the next step will be to broaden the scope of MH by opening the platform to MSs and interested organisations and enabling them to customise the tool for their own purposes. This should contribute to a reduction in the reporting burden for MSs.

MH has already been in use for more than 2 years and supports the implementation of standards at the ESS level (with the collection and sharing of reference metadata files for approximately 60 statistical domains) and at national level (used by National Statistical Institutes). Optionally, at the level of European Commission, MH could be used by other DGs to create quality reports and other descriptive metadata for the statistics that they produce.

3.3. Harmonisation process and its logical steps

Eurostat has identified 104 statistical production processes[5]. For 80 of these, there are national data collections/ data transmissions, which therefore require national reference metadata. So far, Eurostat and countries have standardised national metadata covering 37 out of these 80 processes. It should be noted that several ESMS reference metadata flows are often opened per statistical process. For instance, 23 metadata flows (one flow per statistical indicator) have been opened for the statistical process "Short-Term Business Statistics". Most of the time, one ESQRS is used per process.

The process of harmonisation starts with **Preparation** – during this phase all the necessary information to identify the needs for descriptive metadata and quality reports of the statistical domain is collected. Furthermore, the identification of the information that is collected from the MSs and the precise format in which that information is collected (word, excel or PDF) is carried out, as well as information about any domain specific regulations regarding frequency and content. If there is no domain specific regulation, then the collection falls under the scope of the Regulation 223 [5] which stipulates the various quality dimensions.

The next step is **Compliance** assessment – during this phase, the information collected in the previous step is used to determine which reporting structure (ESMS/ESQRS) is best suited. The ESMS structure is preferred when the information has a general character and for the purpose of dissemination to the general public. The ESQRS structure is preferred when more detailed information is needed and when the information is not directly disseminated except in the case of creation of an aggregated quality report. During this step, each collection is mapped to the relevant standard structure without losing information. Areas for improvement in the report are also identified. As this step is the core business in the implementation of both

[5] A statistical production process is defined as the collection, processing, compilation and dissemination of statistics for a defined area and with a specified frequency.

standards, it is logically the most time consuming and requires a constant dialogue between Eurostat and MSs until common agreement is reached.

The third step is **Implementation** – this is the phase in which all the available metadata are transferred to the standard IT platform (the Metadata Handler in the case of Eurostat) using the standard reporting structures. This phase ends with the confirmation from the MSs experts that all the information was preserved. This is facilitated by the tool which provides the workflow approval functionality (Draft > Ready for Validation > Validated).

The last step is **Production** – in this last phase, the Reference Metadata files (ESMS/ESQRS) are stored in the Metadata Handler, a reference standard environment for quality monitoring at national and Eurostat level. This information can also be disseminated to the general public if agreed by MSs during their domain specific Working Groups.

Metadata files then enter the cycle of regular revisions according to each statistical domain data collection frequency.

In concrete terms, there are 888 ESMS files currently disseminated by the MH. Of these, 612 are national ESMS files produced by Member States and other participating countries (47 countries in total) and 276 are European ESMS files produced by Eurostat.

4. Conclusions and the way forward

Standardisation of structural, reference and other metadata and making them available to general public is becoming more urgent with quickly raising demands for automatized machine to machine data and metadata transmission. Indeed, efficient data exchange, validation and dissemination is not possible without putting into place appropriate statistical data models and standards. In this context, Eurostat plays an important coordinating role both at the level of European Statistical System (ESS) and at the global level within the context of collaboration with the United Nations. Eurostat is one of seven sponsoring organisations of the SDMX initiative to foster standards for the exchange of statistical information. Future practical implementation of standards by using appropriate data and information models (such as Data Structure Definitions defined by SDMX Information Model) must inevitably be accompanied by the continued development of supporting IT tools, reference repositories, guidelines and training manuals. These tools and standards must be developed in a way which will enable them to be adapted and reused by a broad range of other organisations at national, ESS and international levels. Strong demand for linking and disseminating publicly-available (open) statistical and non-statistical data will not be satisfied without well-structured and harmonised metadata.

References

[1] A. Bikauskaite, L. Gramaglia, A. Gotzfried, and H. Linden (June 2014), Better data quality through global data and metadata sharing, Vienna, Q2014.

[2] Common Metadata Framework, METIS (2013), UNECE http://www1.unece.org/stat/platform/display/metis/The+Common+Metadata+Framework

[3] ESS Vision 2020, Building the future of European statistics, ESS, (2015) European Union http://ec.europa.eu/eurostat/documents/42577/6906243/ESS+vision+2020+brochure/4baffcaa-9469-4372-b1ea-40784ca1db62

[4] Eurostat internal paper (February 2016), Coordination of the ESS Vision 2020 with the activities of the UNECE High Level Group for the Modernisation of Official Statistics, Eurostat internal paper.

[5] Regulation (EC) No 223/2009 of The European Parliament and of The Council on European statistics, (March 2009) http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32009R0223&from=EN

[6] Technical Manual of the Single Integrated Metadata Structure (SIMS), (2014), Eurostat http://ec.europa.eu/eurostat/documents/64157/4373903/03-Single-Integrated-Metadata-Structure-and-its-Technical-Manual.pdf/6013a162-e8e2-4a8a-8219-83e3318cbb39