Distr. GENERAL
29 April 2013
WP.5
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN UNION (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION
AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

Work Session on Statistical Metadata
(Geneva, Switzerland, 6-8 May 2013)
Topic (i): Metadata standards and models

DDI-SDMX INTEGRATION AND IMPLEMENTATION

Working Paper

Prepared by Marco Pellegrino and Denis Grofils, Eurostat

I. Introduction

1. The European Statistical System's Vision Infrastructure Programme (ESS.VIP) seeks to achieve the overall goal of the Commission Communication (COM (2009) 404) on the Production method of EU statistics: A Vision for the next decade and to put in place a more integrated production system of European statistics founded on common structure, subsidiarity and minimum standards for using shared elements of the system.

2. The ESS VIP Programme is considered as an indispensable part of the European Statistical Programme (ESP) 2013-2017: appropriate planning has to ensure a fruitful interaction with other parts of the ESP 2013-17 (e.g. the modernisation of statistical domains, development of indicators for well-being, etc.). It takes fully into account the outcomes from the work of the High Level Group (HLG) on modernisation of statistical production set up by the CES Bureau in 2010.

3. The basic idea of the Programme is to realise economies of scale and productivity gains through sharing data, services and costs and through better allocation of tasks among the ESS partners. Only in this way, ESS partners will be able to meet growing challenges under budgetary constraints. In November 2012, the ESS Committee supported the overall goals of the ESS VIP Programme and underlined its strategic importance for the future of the ESS. While the first three projects of the Programme (SIMSTAT, Common Data Validation Policy and European System of Statistical Business Registers) were endorsed, others are being elaborated.

II. A cross-cutting project on Information Models and Standards

4. Within the ESS VIP programme, a cross-cutting project on Information Models and Standards (CRC.IMS) aims at defining information standards at ESS level, ensuring consistency across VIP projects and providing - where relevant - technical solutions for mapping project activities to existing relevant standards. Through these activities, the ESS also contributes to the development of

international standards in the relevant areas, in coordination with work groups at UNECE (GSIM, Plug & Play) and with the HLG Modernisation Committee Standards (now including the DDI-SDMX dialogue), trying to achieve a broader consensus.

5. The Sponsorship on Standardisation strategic task force recently highlighted that "the European Statistical System has a long tradition in harmonising statistical products and regulating requirements within the different statistical domains. International cooperation has not put much effort on harmonizing production methods, processes and systems, however. In 2010, the highest ESS governance body, the ESSC, adopted a Joint Strategy. This Joint Strategy envisages further integration of the ESS and states – among other things – that this will require more harmonisation and standardisation of statistical methods, of the IT infrastructure and IT tools, and of metadata. All of this will eventually lead to better quality and higher productivity of the statistical data processing".

A. The current state

6. The landscape of statistical standards is evolving and the maintenance of a relevant portfolio of standards is more than ever a critical activity for statistical systems. Besides well-established ESS standards – as the Statistical Data and Metadata eXchange[1] (SDMX) – other initiatives are gaining importance in the statistical world, such as:

    - The Data Documentation Initiative[2] (DDI) originating for the world of social sciences data archives and more and more in use in statistical organizations for the documentation of micro-data.

    - The Generic Statistical Business Process Model[3] (GSBPM) providing a common abstract representation of statistical activities

    - The Generic Statistical Information Model[4] (GSIM), a common abstract representation of data objects manipulated in official statistical production and elaborated as an overarching model for implementation standards such as SDMX or DDI.

7. GSIM, in particular, is seen as one of the cornerstones for modernizing official statistics and moving away from subject-matter statistical stovepipes. As such, it is also part of the strategic vision prepared by the High-Level Group for the Modernization of Statistical Production and Services (HLG). GSIM provides a standardised statistical language for describing information objects, which are the inputs and outputs in the production of statistics. As a reference framework, GSIM helps to explain significant relationships among the entities involved in statistical production, and can be used as an overarching information model allowing the integration of the other standards (DDI and SDMX among the others).

8. Nevertheless, after years of discussions in international working groups, the current situation is far from being optimised:

    - Most of the ESS data are described without a standard information model. In most cases, data are described either directly in regulations or through Excel templates, without explicit reference to a common set of rules. Cross-domain standardisation of concepts and codes is the exception.

    - In data transmission, formats such as CSV (comma separated value) are used, where the column descriptors are defined in separate message implementation guides. Different versions of GESMES are used (where only GESMES/TS is part of the SDMX standard) with data descriptors (*key families*) stored in separate databases, and Genedi tools used for validation.

[1] http://sdmx.org
[2] http://www.ddialliance.org
[3] http://www1.unece.org/stat/platform/display/metis/The+Generic+Statistical+Business+Process+Model
[4] http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=59703371

- In some cases, SDMX is used for designing so-called "structural" metadata. SDMX (ISO International Standard 17369) is designed to describe statistical data and to normalise their exchange. Its use is a business choice aiming at a reduction of development, maintenance and operation costs for an organisation. But SDMX is focused on tabular aggregates, with some preference for time-series and relatively small data sets, and its use for micro-data management is still under discussion.

- There is no commonly agreed standard supporting other business needs, such as frequent exchanges of transaction data, micro-data exchanges and sharing, interconnection and automation of services.

B. Scope of the IMS project and objectives

9. The "Information Models and Standards" project intends to ensure that the European Statistical System has a pro-active role in the choice and fine-tuning of a set of standards and models supporting the modernisation of statistical services. These standards and models are in some cases already in use, or in advanced elaboration phase, and will be further enhanced and maintained by the international community of official statistics.

10. The METIS Task-Force on a Common Metadata Framework has already published in 2010 a draft list of standards, models, best practices and other methodological materials available on the web, while the High-Level Group on Modernisation of Statistics (HLG) and the Sponsorship on Standardisation task-force promoted a new inventory of standards used within the ESS. The Cross-Cutting project on Information Models and Standards will take these contributions into account, trying to foster a further integration and enhancement of some relevant technical and statistical standards, of the IT infrastructure, and of metadata management within the ESS.

11. In cooperation with the ESS statistical institutes, our objectives are:

    - To ensure that the European Statistical System (ESS) has access to a set of agreed-upon standards supporting the modernisation of statistical services, so that the ESS may also play a leading role in the evolution of these standards.

    - To increase coherence between standards, at the same time ensuring that these are consistent with best practices and recommendations from the international community of official statistics.

    - To define information models that can be used across the ESS to model structural metadata for different types of data, taking into account existing standards and on-going developments in the statistical world.

    - To identify and adopt guidelines for documenting business processes in the ESS.

    - To provide support mechanisms (e.g., capacity-building and training) for the practical implementation of these standards and models.

12. The "Information Models and Standards" is a 2-year project which consists of 5 phases:

    - PHASE 1: Project setup & interaction with the VIP programme (already started)

    - PHASE 2: Analysis of standards and models (starting)

    - PHASE 3: Proposal for an integrated standard to handle micro-data and aggregated data

    - PHASE 4: Maintenance and further enhancement of standards

    - PHASE 5: Capacity-building (training, implementation consultancy, and support to ESS users).

13. An initial project phase assuring the integration in the overall ESS VIP programme – through identification and analysis of interdependencies with other cross-cutting projects – has started. A set of activities linked to the analysis and evaluation of various standards is starting with a view of delivering a first cut of a proposal for an integrated micro-macro framework based on SDMX and DDI at the end of 2013.

14. Note that this project is not an IT project, although it has an important IT dimension. This is, first of all, a methodological development that is an enabler for establishing the required IT solutions. The current project will deliver methodological standards and implementation guidelines, which in some cases will need specific activities for the adaptation of the IT infrastructure.

III. An integrated standard to handle micro-data and aggregated data

15. The use of standards presents several clear advantages:

- It enables a standardised description of all statistical data structures within the scope of the model, enforcing the use of standardised concepts and codes whenever possible.

- It enables the integration of data and metadata, making it easier to locate, retrieve, use and understand statistics in all statistical domains.

- It speeds up the exchange of data, increasing efficiency and quality, and reducing management costs.

- It enables the design and re-use of generic software services for collection, processing and dissemination (e.g., web services, data validation, visualisation and automated loading of data and metadata, querying).

- It helps mapping and interconnecting specific database models used by services and applications located in different statistical systems.

16. Given the number of different initiatives dealing with standards and models for micro-data and aggregates, time seems to be mature for the ESS to come up with a clear and coordinated approach on which standard should be used – and for doing what – in the field of micro-data management. In the absence of clear guidelines, the risk of creating stovepipes and diverging developments in different organisations is very high, and with this the consequence of creating statistical and IT frameworks which are inefficient and not interoperable. The choice is not black-and-white: both DDI and SDMX have specific capabilities which may or may not be relevant to the implementer, depending on the specific use case. There are also implications for how useful different types of tools based on the standards are, which should not be ignored.

17. An evaluation of the two standards with regard to microdata can only be made in the context of precise use cases. If we are faced with the straightforward retrieval of simple sets of administrative data, for instance, we may not need the extra metadata capabilities provided by DDI. If data are more complex or we need more exhaustive metadata, or we are describing data collected by a sample survey or census, on the other hand, DDI is better suited for this purpose. Each data collection and production scenario has its own demands as regards data and metadata, and either standard may be better suited, depending on those requirements.

18. The current state of discussion at technical level (within the DDI-SDMX dialogue and the SDMX Technical Working Group) has elaborated the following categories of use cases for which the use of DDI and SDMX can be assessed:

- ✓ **Survey data collection:** Survey-based data collection requires the structured representation of questionnaires and software tools capable of manipulating the questionnaire representations and instantiate them into survey data. DDI offers built-in schemas for the representation of

questionnaires. Different tools support the development, usage and export of questionnaires in DDI format. SDMX, on the other hand, does not offer natively metadata schemas but offers a facility for user communities to define their own schemas (as it is done for example with ESQRS in the area of quality reporting). Still, no widely accepted SDMX schema exists for questionnaires and, as a consequence, no widely used tool exists to support survey-based data collection in SDMX format. The Labour force survey (LFS) will support this use case.

- ✓ **Administrative & register data:** Register data often represent high-volumes of data which could constitute an obstacle to its representation in XML formats. The maintenance of statistical registers generally implies transactional processing for which a consistent modelling is needed. The Euro-Groups Register (EGR) and the Banca d'Italia Debt Securities Register will support this use case. Standard solutions for exchanging high volumes of administrative or register data as well the consistent representation of transactions on registers will be tackled.

- ✓ **Reference environment for the combined use of DDI and SDMX:** The overall objective of this use case is the setup of an integrated environment for handling micro-data (from surveys and administrative sources) and further aggregations (statistical tables) using the appropriate standard and taking advantage of a set of cross-domain content guidelines and software tools. A generic integrated process would then enable to use DDI for describing, documenting and driving data collection, editing, imputation and estimation, whilst SDMX is used for further aggregation, dissemination and exchange using web services and a standard SDMX Reference Infrastructure. The project will assess the feasibility of a standard-agnostic metadata registry which could manage metadata compliant with different standards, such as (but not only) SDMX and DDI.

- ✓ **Statistical micro-data access and on-demand tabulation of micro-data:** Eurostat is already engaged in providing access to micro-data sets for scientific use within its secure data centre. The data sets are important from the research perspective, as they are harmonised across Europe. This use case involves exploring whether the documentation and access to micro-data made available from Eurostat for research purposes would be improved by the use of DDI. Using DDI to support data management and documentation for researchers is a very common use case. This project could use the data being marked up within the Data without Boundaries (DwB) project. The output from this would be a set of recommendations on how Eurostat might deploy DDI in support of its secured data centre.

- ✓ **Metadata and quality reporting:** Eurostat uses SDMX-based formats for metadata and quality reporting (ESMS and ESQRS). Other customised formats are still used for collecting from NSIs information about the data and their quality which then must be channelled through ESMS and ESQRS. Some agencies which are quite advanced in using DDI (notably INEGI in Mexico) have experimented with populating ESMS and ESQRS metadata reports with the contents of already existing DDI metadata repositories. This use case would involve exploring how metadata collection and assessment might be improved for Eurostat by employing DDI in some capacity. This also depends on the possibility of deploying DDI repositories in Member States. The output, at this stage, would be a set of recommendations for the joint use of DDI and SDMX for metadata collection and exchange within the ESS.

IV. Conclusion

19. A consistent coverage of the statistical production process by appropriate standards across the European Statistical System is a key element for achieving the ambitions of the ESS VIP programme.

20. There is a clear interaction between this project and the "Framework and Standards for statistical Modernisations" recently discussed by the High-Level Group for the modernisation of statistical production and services (HLG), in particular with the work tasks dealing with GSIM and its mapping to DDI and SDMX. A DDI-SDMX integrated production flow implies transitions from one standard to the other. For example, a typical case is DDI-documented statistical micro-data sets which can be further aggregated and exposed as SDMX datasets. Various other cases can be considered, such as

processing inputs in DDI and SDMX and outputting in one or another standard. The integrated use of DDI and SDMX requires concrete mappings between DDI and SDMX constructs, and some choices about how data elements are handed over from one model to the other. In this respects, the development and implementation of GSIM as a common abstract model, and mapping of GSIM to DDI and DMX, is a key step in this direction. The DDI-SDMX integration activity aims at assigning one or both standards to various possible production steps, possibly per data type (survey data, administrative data, register data) and provides a model for the transition from one standard to the other. In the end, ESS statistical institutes should know clearly which standard to use for which data and how to pursue their processing process at architectural and implementation level.

21. The cross-cutting project on Information Models and Standards (IMS) is part of the ESS VIP programme and, as such, will be conducted in coordination with other VIP projects and in cooperation with member States, making use of a well-established governance framework based on thematic Directors' Groups, task-forces and working groups.