SDMX as the logical foundation of the data and metadata model at the ECB

Gérard Salou[1] and Xavier Sosnovsky[2]

Introduction

The Directorate General Statistics of the European Central Bank (ECB) is responsible for efficient management of the statistics that are needed for the ECB's monetary policy and the other functions of the ECB, the Eurosystem and the European System of Central Banks (ESCB). In addition, it is responsible for providing statistics to the interested public and market participants. For those tasks, the ECB has developed a statistical infrastructure based entirely on SDMX[3] standards. This paper starts with a description of the SDMX statistical standards, covering the development of the standards, the underlying data model and implied data organisation. Later, the paper describes how SDMX standards are used as the common logical foundation for data collection, data dissemination and data models within the ESCB. Tools and applications of common interest to the SDMX community are also described, in particular the ECB SDMX loader suite, which is used by the ECB for converting seamlessly between data formats (EDI, XML, CSV), the ECB data visualisation tools and the related ECB SDMX framework.

In a second section, the paper describes the SDMX service-oriented architecture developed by the ECB. The web services allow the retrieval of data in the SDMX-ML formats based on filters such as dimensions, attributes, time, datasets, data flows and categories. They allow users to retrieve the latest version of the data, specific snapshots, as well as updates and revisions. They also allow the retrieval of structural metadata such as category schemes, dataflow definitions, data structure definitions, concepts, code lists, as well as "maintenance agencies", data providers and organisation schemes. Two complementary technical implementations of the web services are also presented. The system is also supplemented with a subscription and notification mechanism.

The paper concludes by describing the benefits in terms of efficiency and effectiveness of using SDMX standards for statistical data and metadata infrastructures. The experience of navigating SDMX-ML data in the same way as HTML pages is presented as a vision for data browsing in the future.

The development of SDMX standards

Since their creation, the ECB and the ESCB have used a standard format for their data interchanges. At the time (1998), the format had been based on the Electronic Data Interchange (EDI) technology, using the EDIFACT syntax. The precise implementation

[1] ECB, Statistical Information Services Division, Kaiserstraße 29, D-60311 Frankfurt am Main, Germany. E-mail: gerard.salou@ecb.europa.eu.

[2] ECB, Statistical Information Services Division, Kaiserstraße 29, D-60311 Frankfurt am Main, Germany. E-mail: xavier.sosnovsky@ecb.europa.eu.

[3] A glossary of acronyms is available in the annex.

chosen by the ESCB, the GESMES/TS message, was derived from a more generic EDIFACT standard called GESMES (Generic Statistical Message) created in the early 1990s by a group of institutions under the leadership and support of Eurostat.

At the beginning of the 2000s, a group of institutions, including the Bank for International Settlements (BIS), the ECB, EUROSTAT, the International Monetary Fund (IMF), the Organisation for Economic Co-operation and Development (OECD), the United Nations and the World Bank, joined together to work on business practices in the field of statistical information with a view to developing more efficient processes for exchange and sharing of data and metadata for their statistical activities. The goal of the consortium was to take advantage of developing technologies and ongoing standardisation activities that could allow them to gain efficiency and avoid duplication of effort in the field of transfer of statistical information.

In that context, the GESMES/TS EDIFACT message, intensively used by the ESCB, the BIS and Eurostat, was very soon thereafter adopted as the first SDMX data format standard under its present name of SDMX-EDI. Based on the same data model, the consortium also developed a data format using XML and web technologies. The data model underlying the original GESMES/TS message was further elaborated upon and eventually a new set of standards was adopted. The present version of the standards is now version 2.0 and has become an ISO technical specification (ISO/TS 17369:2005).

The SDMX standards cover[4]:

- The SDMX information model underlying all other elements of the standards;
- SDMX-EDI: the EDIFACT format for exchange of SDMX-structured data and metadata;
- SDMX-ML: the XML format for exchange of SDMX-structured data and metadata. This is accompanied by a set of XML schemas (for example, the schemas for the generic, compact, utility and cross-section data formats);
- The SDMX Registry Specification: this specification defines the basic services offered by the SDMX Registry: registration of data and metadata, querying for data and metadata, and subscription/notification regarding updates to the registry;
- The SDMX Guide – a guide to facilitate the use of the SDMX specifications. It includes reference material for the use of the SDMX information model and provides some best practices for assigning identifiers and designing data structure definitions;
- Web Services Guidelines: a guide for the implementation of SDMX using web services technologies. It places emphasis on web services technologies which will work regardless of the development environment or platform used to create the web services.

The SDMX information model

SDMX standards are based on a powerful information model which is an elaboration on a star schema in which datasets are represented as multidimensional cubes, as commonly done in the information technology (IT) industry for data warehousing. The data structures are described by the so-called structural metadata. These include definitions about the dimensions and the related attributes that are relevant and used in the context of a dataset.

[4] For further details, see www.sdmx.org.

Structural metadata comprise the statistical concepts, code lists and data structure definitions (DSDs) which provide the basis for the data representation of each dataset. Each DSD provides definitions and additional details for the dimensions and attributes used in representing data in a particular subject-matter domain. Statistical concepts such as frequency, reference area and institutional sector can play the role of dimensions. Concepts such as units of measurement or methodological comment can be attributes. While concrete dimension values are used to identify and point to a concrete data point (or set of data points), concrete attribute values provide qualitative information about observations (or sets of observations). In particular datasets, dimensions take their values from predefined code lists. Attributes, on the other hand, can take their values from code lists (coded attributes, eg unit of measurement) or can be free text (non-coded attributes, eg observation status). In practice, some attributes must always have a value (they are defined as mandatory in a DSD) while others only have a value for each of the object(s) they are associated with under certain conditions (conditional attributes).

In general, structural metadata are created, administered, maintained and disseminated by the institution managing the data. In the case of the ESCB data, the Statistical Information Services Division of the ECB's Directorate General Statistics coordinates and performs these activities. This enforces discipline in data management, in particular for datasets used across the ESCB. Furthermore, the standardisation work on the data representation also facilitates standardisation in the higher level layers (eg code lists), which yields benefits for the statistical community at large that become increasingly evident. There are cases where institutions use structural metadata defined by other institutions for their own datasets (eg the more global harmonisation is achieved in a particular statistical subject matter, the more likely it is that the structural metadata used in this domain are managed via a collective agreement or by the organisation coordinating the particular harmonisation effort). SDMX facilitates development in that direction. The following section elaborates on the role of DSDs in data management.

The importance of the DSDs in the information model

The set of data managed by an institution can be partitioned into a number of datasets in such a way that each dataset can be represented by no more than one DSD. A DSD can be used for several datasets. In reality, there are several ways to partition datasets and several ways to define alternative DSDs for a given dataset. It is generally desirable to avoid having too many datasets and too many DSDs because it increases maintenance costs. It is also important to partition data in such a way that DSDs share concepts so that datasets can be more easily related using common dimensions. Research has been done in that domain, in particular by Sundgren (2006) with the so called "αβγτ" model. A particular application to SDMX is developed by Androvitsaneas et al (2006). For example, a large macroeconomic database could be partitioned into national and financial accounts, prices and short-term statistics, external sector statistics, government finance statistics and monetary and banking statistics. A DSD could be specified for each one of these blocks, and if there is a need, for example, for more efficient data management, each block could be partitioned into subsets using the same DSD.

Moreover, a more detailed breakdown could be envisaged at a higher level if there is a need for more flexibility and room for manoeuvre in the specification of DSDs, thus foreseeing a different DSD for each of the more detailed categories/partitions (eg splitting monetary and banking statistics into monetary statistics and banking statistics and specifying a corresponding DSD for each one). Actually, the more aggregated the data are, the smaller is the need for partitions and DSDs (as in the case of dissemination from central supranational institutions). The more detailed the data that are managed are (as in the case of data collected by a national statistical institute), the higher is the number of partitions and DSDs required.

The ECB has defined and constantly maintains more than 40 DSDs and a much higher number of partitions or data flows (several data flows may use the same DSD). Most of these DSDs are also used in the data sharing process within the ESCB and with Eurostat, thus contributing to harmonisation not only at the technical level but also in the statistical modelling layers, up to the content. The ECB also uses several DSDs defined and managed by other institutions disseminating data in SDMX formats, mainly the BIS and Eurostat, which are applicable when ECB end-users access data from those other institutions. Similarly, the ECB internally creates "artificial" DSDs for accommodating, within its internal statistical data warehouse data model, data from other sources, such as the OECD or the IMF, that do not yet disseminate DSDs for all their datasets. It is interesting that in some areas (eg balance of payments) the lead role in defining DSDs is played by one institution (IMF), while other institutions (ECB, Eurostat, OECD) may also use them, directly or indirectly, in their data management work. The maintenance of DSDs is also of very high importance, since they form the basis of data content and also constitute the link to other data structures which may exist for related data. For example, special care is taken with the structural metadata used in the domain of balance of payments statistics (BOP), for which the ongoing world-wide use of SDMX-EDI requires a high level of coordination among the institutions involved in administering BOP data exchanges (ECB, Eurostat, IMF, OECD). In conclusion, the use of structural metadata is of crucial importance for promoting harmonisation and maximising efficiency in statistical activities, since it makes data structures more accessible and visible to non-IT specialists.

SDMX along the statistical process

The ESCB having adopted the SDMX standards for all its data exchanges, it was very natural for the ECB to also consider the SDMX data model for adoption and use all along its statistical processes, from data collection to data dissemination. The data model was implemented in the internal database system used for statistical compilation and subsequently in the ECB Statistical Data Warehouse. Later, the use of the SDMX model and format was expanded to the data dissemination website.

To assist the statistical processes, a number of SDMX-based tools have been developed by the ECB. These include the ECB loader suite, which provides data loaders and writers as well as a data checker that verifies the compliance of incoming files with the standard format. That tool also provides some format translators, to convert from one representation to another, covering SDMX variants and standard IT formats such as CSV.

On the dissemination side, a number of implementations can be seen on the ECB and NCB websites. In addition to the fact that SDMX-ML files are made available throughout the statistics section of the ECB website, the most technically interesting case is the Eurosystem Joint Dissemination, where euro area data and their respective national contributions are updated on the ECB website and replicated automatically on the websites of national central banks (NCBs), using the national language(s) and the look and feel of the respective NCB website. This is made possible through the use of web technologies (XML and XSLT) in combination with SDMX standards.

The usage of the SDMX data model for the ECB Statistical Data Warehouse (SDW) was also a natural choice, given the fact that the underlying data model is close to a star schema which itself is commonly used for data warehousing in the IT industry. Using those technologies and sound DSD management makes it possible to relate data from different datasets and to combine them into virtual datasets using their common dimensions. It then becomes possible to present data not from the producers' point of view, as is the case with the original DSD, but rather from the users' point of view, using virtual datasets.

Finally, in order to improve the visual display and the accessibility of data, and to make data analysis more efficient, productive and successful, the ECB has created various visualisation tools on the ECB website. SDMX is at the core of these rich internet applications. Not only do they consume data expressed in SDMX-ML, but they are also modelled according to the SDMX information model. The applications have become very popular and they have definitely improved the understanding of the statistics published.

Implementation of SDMX web services

In order to allow interested parties to use data available in the SDW in their own applications or keep their databases automatically up to date, the ECB implemented a set of web services which follow the SDMX specifications. The web service makes it possible for all potential clients to interface their applications with the SDW and thus have online access to data with comprehensive euro area, national and international data coverage. It also facilitates the reproduction of ECB statistics on other media by interested parties such as information distributors. Furthermore, the web service makes it easier for international organisations, such as the OECD and the IMF, to feed their own data systems with ECB statistics. The following paragraphs describe the SDMX web services.

As defined by the World Wide Web Consortium (W3C), a web service is a software system designed to support interoperable machine-to-machine interaction over a network. A web service involves APIs (application programming interface) that can be accessed over a network, such as the internet, and executed on a remote system hosting the requested services. Web services exchange data in XML format. Using SDMX, the web service inherits from:

- The SDMX information model, for describing statistical data and metadata;
- A proposed API for web services;
- The SDMX-ML query format, a standard for requesting data. The query message supports the retrieval of both statistical data and metadata, using various filters;
- Various formats in SDMX-ML for supplying data (eg SDMX-ML generic data, SDMX-ML compact data and SDMX-ML structure formats).

Additionally, these SDMX standards build on general IT standards such as HTTP, XML, SOAP and WSDL, or REST. The combination of these IT standards with the SDMX statistical standards makes it possible for the data user to abstract from the IT hardware, network, database and language layers used on the data side.

The SDW web service is designed to:

- Offer access to the statistical data stored in the SDW, including the possibility of obtaining updates and revisions only, or snapshots at specific points in time. The data can be filtered by data flows, datasets, dimensions, attributes, date ranges, etc. The matching data can be returned in the SDMX-ML generic data or SDMX-ML compact data formats;
- Offer access to metadata information such as data structure definitions (DSDs), concepts, code lists, dimensions and economic concepts, so that the statistical data mentioned above can be easily understood, used in calculations, automatically processed and interpreted. The matching metadata will be returned in the SDMX-ML structure format;

- Offer a subscription/notification mechanism whereby users of the web services receive a notification when data of interest have been updated.

This approach follows current best practices, as the “publish once and pull” scenario underlying the implementation of SDMX web services makes it possible to take full advantage of the data sharing model in facilitating low cost, high quality statistical data and metadata exchange. If properly deployed, an SDMX-based architecture coupled with an SDMX web service allow for a very high degree of automation for data collection, dissemination and processing for the benefit of the wider statistical community and beyond (institutional policymakers, researchers, and others).

Conclusions

This paper has described how the early adoption of the SDMX statistical standards by the ECB, and its constant efforts to keep its systems compliant with the standard, has made it possible for the ECB to maintain its statistical systems at the leading edge of technology, from data collection to data dissemination. For statistical institutions the main benefits provided by the implementation of the SDMX standards all along the statistical process are as follows:

- Full harmonisation of data representation for ESCB data sets;
- Fully automatic data exchanges within the ESCB;
- Unique statistical data warehouse for aggregated data within the ECB;
- Reusability of components and methods for IT tools along the process;
- More efficient development of tools due to improved interoperability and increasing abstraction of IT layers;
- Synergies with technological developments.

In addition, because SDMX includes a well-developed and tested data model together with a collection of techniques and tools, the development of a modern statistical system based on SDMX is simpler, cheaper and more efficient.

As shown by the implementation at the ECB and on its website, it is possible to develop applications that recognise the various statistical objects contained in SDMX-ML files. This means that it is possible to create applications that can read and understand statistical data independently of their location, origin and IT implementation, in the same way as a web browser can interpret and use HTML pages or a music player interprets MP3 files. With this analogy it is possible to imagine how the generalised use of the SDMX standards by producers of official statistics could change the way that decision-makers and citizens could access and use statistics in the future. They will no longer need to use data interfaces specific to the data source, for example, the ECB SDW interface for ECB data, but can use the interface of their choice for all SDMX-enabled sources. Users will easily be able to use data from different sources at the same time and bookmark, refresh and see data as tables, graphs or maps.

Annex:
Glossary of terms

API	Application programming interface: a set of routines, data structures, object classes or protocols provided by libraries in order to support the building of applications
DSD	Data structure definition (also known as “key family”)
EDI	Electronic data interchange
EDIFACT	Electronic Data Interchange for Administration, Commerce and Trade, the international electronic data interchange standard developed under the United Nations
ESCB	European System of Central Banks (27 European Union central banks plus the ECB)
GESMES	Generic statistical message, implemented using the EDIFACT syntax
ISO	International Organization for Standardization
NCB	National central bank
REST	Simple interface which transmits domain-specific data over HTTP without an additional messaging layer such as SOAP
SDMX	Statistical data and metadata exchange
SDW	ECB Statistical Data Warehouse
SOAP	Based on XML, SOAP defines an envelope format and various rules for describing its contents.
URL	Uniform resource locator
W3C	The World Wide Web Consortium is the main international standards organization for the internet (www.w3.org).
WSDL	Web service description language
XML	Extensible mark-up language
XSLT	Extensible stylesheet language transformations (XSLT) is an XML-based language used for the transformation of XML documents.

References

Androvitsaneas, C, B Sundgren and L Thygesen (2006): “Towards an SDMX user guide: exchange of statistical data and metadata between different systems, national and international”, presented at the Meeting of the OECD Expert Group on Statistical Data and Metadata Exchange, Geneva, 6 and 7 April.

Sundgren, B (2006): “Reality as a statistical construction – helping users find statistics relevant for them”, presented at the European Conference on Quality in Survey Statistics (Q2006), Cardiff.