

SDMX and the statistical production process

Gabriele Becker ¹

Summary

SDMX, the Statistical Data and Metadata Exchange initiative sponsored by 7 international organisations, started out to provide technical standards and content-oriented guidelines to improve mechanisms and processes for information exchanges between statistical organisations, both at national and international level. It also aimed at enhancing the web dissemination of statistical data. A growing number of statistical organisations, as well as central banks, are now clearly adopting SDMX technical standards and guidelines in their work. As a result we can observe generic implementations of the SDMX standards framework within a statistical organisation and see the related impact on the efficiency of statistical processes. This paper describes the linkages between SDMX and the statistical production process and points out potential efficiency gains.

1. The statistical production process

The ultimate aim of any official statistical production process is to compile correct and meaningful numeric information to be used by administrations, policy makers, researchers and also, the general public. At the national level there are usually several organisations concerned, national statistical institutes (NSIs), central banks (CBs) and possibly specialised administrations, e.g. for labour statistics. At the international level there are many organisations involved, last but not least the seven SDMX sponsors.

1.1. Stove pipes

The compilation of statistics for a specific domain is usually considered, by those who do it, a highly specialised activity, requiring IT systems and processes customised to this set of statistical data. As a result the statistical production process in many organisations is separated according to the statistical domain, with each one working within its own „stove pipe“. This has often lead to a lack of harmonisation within and also across organisations with respect to how data is organised, what metadata is provided with it to make it meaningful and, in particular, how it is exchanged between „stove pipes“ or between organisations. Even within a statistical organisation it can be difficult to share, for example, IT applications across different subject matter domains, thus creating the potential for inefficiency and duplication of effort. For the end users this means that it is often difficult to use statistical information on different subjects or from different providers in an efficient way.

1.2. Statistical production process: a generalisation

The Joint UNECE / Eurostat / OECD Work Sessions on Statistical Metadata (METIS) have over the past years worked on a Generic Statistical Business Process Model (GSBPM) and the possible relationship with SDMX has been recognised: “The GSBPM should therefore be seen as a flexible tool to describe and define the set of business processes needed to produce official statistics. The use of this model can also be envisaged in other separate, but often related contexts such as harmonizing statistical computing infrastructures, facilitating the sharing of software components, in the Statistical Data and Metadata eXchange (SDMX) User Guide explaining the use of SDMX in a statistical organisation, and providing a framework for process quality assessment.”²

Level 1 of the GSBPM has the following phases:

- Specify needs
- Design
- Build

¹ Gabriele Becker, Head of Statistical Information Systems, Monetary & Economic Department, Bank for International Settlements (BIS). The views expressed are those of the author and not necessarily those of the BIS.

² GSBPM, Version 3.1 December 2008 (<http://www1.unece.org/stat/platform/display/metis/METIS-wiki>)

- Collect
- Process
- Analyse
- Disseminate
- Archive
- Evaluate [not discussed in this paper]

It also contains eight over-arching statistical processes:

- Quality management
- Metadata management
- Statistical framework management
- Statistical programme management
- Knowledge management
- Data management
- Provider management
- Customer management

After briefly presenting the SDMX Framework, we will review the phases with respect to how SDMX could support them, in particular in the context of a generalised IT infrastructure based on SDMX principles.

2. The SDMX Framework

The “Statistical Data and Metadata Exchange” Initiative (SDMX) aims to “develop and use more efficient processes for exchange and sharing of statistical data and metadata among international organisations and their member countries. To achieve this goal, SDMX provides standard formats for data and metadata, together with content guidelines and an IT architecture for exchange of data and metadata. Organisations are free to make use of whichever elements of SDMX are most appropriate in a given case.”³ The SDMX framework consists of:

SDMX Technical Specifications and related tools:

- SDMX Information Model (SDMX IM)
- Two data exchange formats to exchange data and metadata (SDMX-EDI and SDMX-ML)
- SDMX Registry specification and web services guidelines
- SDMX tools

In addition, the SDMX framework provides Content-oriented Guidelines covering:

- Cross-Domain Concepts (definitions and recommended code lists)
- Metadata Common Vocabulary
- Subject Matter Domain list

The technical specifications concentrate on data exchange and data sharing processes. The SDMX Information model (IM) covers the information (statistical data and metadata) that may be exchanged between statistical agencies and also the related process flows, such as information on data provisioning, eg, which agency should report which data at what time. Obviously, information that is exchanged will also have to be stored in statistical systems and a growing number of implementations use it as a model for this purpose and we will elaborate on this in chapter 3.3

The two data exchange formats can be used to actually “package” data and metadata into data exchange messages covering also structural information that helps a receiving application to interpret and (automatically) process the information. Freely available SDMX tools have been developed with a view to being used as demonstration tools for teaching and learning about SDMX, eg building SDMX data and metadata structures or running exchange format transformations. The SDMX Registry specification provides a central registry of available data and reference metadata and a repository for provisioning information, thus constituting the focal point for process automation.

³ SDMX User Guide 2009.1(http://sdmx.org/?page_id=38), p 11.

The Content-oriented Guidelines focus on the harmonization of specific concepts and terminology that are common to a large number of statistical domains. Such harmonisation complements the potential efficiency gains to be achieved when applying the SDMX technical specifications.

3. The statistical production process and SDMX

In this chapter we will look at the different steps of the statistical business process to identify how the application of SDMX technical standards and content-oriented guidelines can add efficiency. The key argument rests on the fact that the standards can be applied across statistical domains leading to a standardisation of the related statistical processes. It should be possible to build a generic processing system that implements these standards, which in turn can then host the data from different statistical domains and their statistical processes.

3.1. Specify need and design

These phases consist of defining all relevant concepts for the statistical activity: what data should be collected and from whom, what information is needed in addition to purely numeric „data“, how (often) will the data be collected, what quality controls need to be performed upon reception of the data, what processing (aggregations, estimations) need to be performed in order to arrive at the final statistical „product“. Who are the clients of this product and how will they get access to the statistical information, what additional information do they need, eg about the collection exercise and the processing, to correctly interpret the information they receive.

Already this short and incomplete description shows that a data collection exercise “is not only about data“, but to a great extent about information about the data, ie “metadata“. This is where the SDMX Information Model (SDMX IM) comes in: its key feature is the extensive model for data and metadata, so that the information “about the data“ needed in the statistical process and by the final users can be represented. Metadata can be attached to the data: statistical data is “identified“ (via statistical concepts, which might be partly chosen from the SDMX Cross-domain concepts) and further “qualified“ by attributes, which can be “free text“ information, eg about the collection methodology for a particular statistic. Metadata can also be attached to other artefacts of the information model, eg to an element of a code list.

Planning and developing the statistical task “using SDMX“ means that the statistical expert responsible for this work needs to take decisions on how to structure the data to be collected, which statistical concepts should identify the data items and what additional attributes (carrying content or also processing information) should be defined. This is actually nothing new and has been done before, the new feature is that SDMX provides a generic model, which can be applied across statistical domains. The model provides a certain rigour and its application fosters the re-use of statistical concepts and code lists.

An organisation’s first such application of the SDMX IM for a new statistical task will require a certain investment into understanding the model and how to use it, however, once the statistical experts gain experience, it will become easier. A key motivation to start applying the SDMX IM for other statistical collection tasks should be the fact that any information modelled according to the SDMX IM can be exchanged using the same standard SDMX formats. A generalised processing environment that “understands SDMX“ or is “SDMX conformant“ will increase efficiency as more statistical tasks are getting integrated.

3.2. Manage metadata

At this point of the argumentation, it may already have become clear that “manage metadata“ in the SDMX view is not a separate step of the statistical process, but an integral part of all steps. Metadata must accompany the data through the statistical process. It can even be argued that basing the statistical process on SDMX can actually make that process “metadata driven“, based on a generic SDMX conformant processing environment.

A key set of metadata used in SDMX are code lists for dimensions (identifying data) and attributes (qualifying data). Managing these centrally and re-using them to the greatest extent possible provides

a natural path towards harmonisation and thus efficiency gains across the organisation as well as better information about the data for statistical analysts and users.

3.3. Build, Collect and Process

For the data collection activity, SDMX technical standards are a natural choice, as they were developed for data exchange. If the data to be collected as part of a statistical task have been structured according to the SDMX IM, then organisations have the choice of two syntaxes for the actual data exchange messages: SDMX-EDI and SDMX-ML.⁴ Being able to use the same exchange format for different data collection activities creates economies of scale for this step of the statistical process. For a reporting agent the fact that the reporting of different data to possibly different statistical agencies can be done in the same reporting format will be an advantage. A statistical agency receiving data for different collection tasks and from different reporting agents in the same format will also achieve efficiency gains.

These arguments lead to the conclusion that it will be beneficial to have generic “SDMX conformant” processing systems that can digest any type of data collection modelled according to the SDMX IM. While SDMX may initially not have been intended to provide a data and metadata storage model, a growing number of organisations are using the SDMX IM to drive the development of their statistical IT systems. We emphasised the importance of metadata for the statistical process and SDMX allows us to exchange metadata together with the data, eg information from the reporting agent about the quality of a specific reported figure, its confidentiality or other attributes that would be important for the receiving organisation or the final end user. The SDMX conformant system would need to be set up in such a way that it can carry this metadata through the actual processing.

Applying an SDMX data structure to a collection exercise will mean that each data item will be clearly identified based on a set of dimensions that make up its identifier or key. This can help in defining (automated) checking routines for incoming data, taking advantage of, for example, an aggregate/component hierarchy on one of the defining dimensions: the hierarchy defines the list of items that would need to be checked against the higher level aggregate. The same holds for a dimension, where “Credit”, “Debit” and “Net” are defined and reported and the relationship between these figures can be exploited to checking purposes.

Large data collection exercises with a large number of providers need to be automated. Via the IM and the Registry SDMX provides support for such automation. It allows to define “Provision agreements” between data providers and a collection agency, that define, which data set (based on a given data structure) should be reported by which providers at what intervals (or specific dates). This can be used to monitor the actual delivery and also to, for example, create warning messages to late reporters.

3.4. Analyse

The statistical analysis as such is not a process that SDMX is concerned with. However, as a prerequisite to any analysis the statistical analysts require “domain intelligence”, which is basically a different word for “metadata”. SDMX has a strong focus on metadata and, if this metadata is carried through the statistical process into the final “data product” offered for analysis, then we can assume that rich metadata will enhance the quality of the analysis. Analysts, who know more about the data they work with, can make better judgements. The data and metadata structures for the data to be analysed should contain attributes into which “domain specific”, “time series specific” or even “observation specific” domain intelligence can be entered by the statistical analyst, if required. It is thus stored together with the data for future reference.

The GSBPM also considers disclosure control as part of the “analyse” phase. As indicated above, the SDMX IM foresees the attachment of attributes at various level, which also includes attributes about the confidentiality status of the item (observation, time series, whole data set) as well as information about the status of an observation, eg whether it is estimated. Applying the same concepts and code lists for this type of “flagging” across the different statistical domain will again provide efficiency gains and add clarity about the data for users.

⁴ For detail on the differences between the two syntaxes, please see Chapter B4 of the SDMX User Guide (http://sdmx.org/?page_id=38)

3.5. Disseminate

SDMX is on its “home turf” when it comes to dissemination of statistical data and metadata. Dissemination comes in different flavours, from the (classic) provision of static tables (HTML, PDF or EXCEL) on the website of a statistical agency, over the exchange of large amounts of data between agencies via fixed file formats, using CDs, DVDs or telecommunication links, to offering end users an interactive User Interface with navigation, query and download facilities. SDMX can facilitate all these means of dissemination.

We start from the assumption, that at this late stage of the statistical process, the data and metadata to be disseminated are already organised according to the SDMX Information Model, in an SDMX conformant database environment. This means that the data is accompanied by its metadata and that the data and metadata structures (together with the required code lists) are also stored in this system. Data extracted from this system in SDMX formats can be “rendered”, together with the accompanying metadata, as HTML, PDF or EXCEL files using automated transformations. Data and metadata extracted in SDMX formats can directly be disseminated on CD or DVD to other organisations.

The full strength of the SDMX approach becomes apparent, when it comes to offering an interactive User Interface (GUI). The metadata that accompany the data according to the SDMX IM is exactly the information required by any “navigation and search” user interface. Typical queries that users will express when searching for data will be “Give me the nominal GDP for the Euro Area, US and Japan”, or “give me the daily and weekly averages 3-month LIBOR rates for the Yen, Pound and US”. Such queries are based on the SDMX data structures, ie the dimensions and attributes for the particular data sets into which this data is modelled. The data structure information together with “constraint” information about which series are actually available in a given dataset, is fully sufficient to build a generic GUI to search in SDMX conformant data sets, which could be, for example, implemented via an SDMX Registry.

The SDMX Registry can actually be seen as the focal point for dissemination activities. The availability of a data set, in a database that can be queried or as a data file on a website, can be registered there. Users of that particular data can subscribe to the registry to receive a notification once that data set has been updated with new information.

3.6. Archive

When archiving data it is important to archive them jointly with the explanatory metadata as otherwise the data will be useless. Again SDMX, due its focus on the combination of data and metadata provides advantages. Data and metadata that have been structured according to a common model, the SDMX IM, are more easily archived than data with different structures. While SDMX does not specifically deal with the issue of archiving, it facilitates it, again due to its generic Information Model and the possibility to package the data and metadata into a common format based on the SDMX standards. Structures and data in SDMX formats (eg in SDMX-ML) are flat files which can be efficiently “zipped” and stored. Together they can be archived and “brought back”, ie loaded again into any statistical environment that “understands” SDMX: the data structure message is used to “interpret” the actual data and metadata message. The data structure message, which is usually a rather small file compared to the data files, contains all relevant information about the content of the (archived) data file, ie all information about the code lists used in the identifying dimensions. As described in chapter 3.5 this information can be used to offer users the option to “search and navigate” also the archived data in the same way as the current data providing them a detailed overview of that is actually available in the archive.

A popular standard among data archives is the Data Documentation Initiative⁵ (DDI). It is worth pointing out that in recent versions of DDI, there has been a conscious alignment between how SDMX and DDI describe aggregate data (unlike SDMX, DDI also focuses on microdata). Thus, it would be possible to transform SDMX data and metadata into the DDI for the purposes of archiving.

⁵ <http://www.icpsr.umich.edu/DDI/>