SDMX from the Dutch perspective: lessons learned

Olav ten Bosch, Roelof Lindeman, Statistics Netherlands

1. Introduction

Launched in 2001, SDMX celebrates its 10th birthday this year. Cause for celebration, but also cause for reflection on the status and expectations of the standard as seen through the eyes of a national statistical institute (NSI). This paper gives an overview of the work on SDMX carried out at Statistics Netherlands in the last few years, and the lessons we have learned.

Starting out with an early participation in the first SODI project, Statistics Netherlands has recently taken further steps into SDMX. After some preliminary exploration of the existing SDMX 2.0 standard and the new 2.1 standard, we have explored some of the available SDMX software with the aim of designing and implementing a general solution for SDMX communication for all statistical domains. We examined and tested two SDMX frameworks in more depth.

As the International Census Hub project moved forward, we decided to use this project as a pilot case. We did some research on technical scenarios to implement the Dutch part of the European Census Hub. Although the decision to actually implement one of these scenarios has still to be made, we have established that multiple scenarios can be implemented. We report briefly about tests using real data from the register-based Dutch Census on one full Census hypercube.

As a result of these activities we have some general ideas about the future of SDMX. We think SDMX is a step in the right direction, but the overall success of the standard heavily depends on how it is used. We summarise our thoughts on this and indicate future challenges from our perspective.

2. Lessons from the first SODI project

In 2005 Statistics Netherlands participated in the International SDMX Open Data Interchange (SODI) working group (also known as the first SODI project). The aim of that project was to make certain short-term national data (i.e. quarterly GDP data and monthly industrial production indices) quickly available in a common dissemination environment of Eurostat via an SDMX web service, based on SDMX 1.0. Statistics Netherlands implemented the web service in a test version of its main output database, StatLine.

At that time our conclusion was that although the original SDMX-ML schemas were quite generic, in practice for each data flow, for the quarterly GDP data as well as for the monthly industrial production indices, a more specialised subset of SDMX-ML with a corresponding specialised web service had to be designed at the NSI [1]. The implication for an NSI in that set-up was the development of a separate web service for each data flow to Eurostat. Although the development of these services could clearly be made more efficient by sharing generic parts, we concluded that this approach was not very efficient from the viewpoint of an NSI, even if we disregard any additional burden as a result of future changes in the SDMX-ML subset designs.

We concluded that, in order to be efficient, generic software designs should facilitate a flexible matching of the output metadata structures of an NSI to the metadata structures referred to in Eurostat requests. We felt that a generic solution to this metadata matching problem was the most burning issue to be tackled for efficient information exchange between Eurostat and NSIs, but this issue has nothing to do with web-service technology in particular.

The demo web service that was implemented within the project was kept operational for quite some time and was used many times for demos on the usability of the concept. Although it never went into production, we feel that the tests carried out in the first SODI project were useful as a first step into the design of a standardised, modern, web-based dissemination infrastructure.

3. A renewed look at the state of SDMX

SDMX moved forward with new versions (2.0, 2.1) and with an international agreement to use the standard for international data communication [2]. Therefore, in 2009 Statistics Netherlands started a project to implement SDMX, in both technical and organisational terms, for communication of statistics to international organisations. To be cost-effective, the approach to be chosen would have to be as generic as possible so that it could be re-used across statistical domains.

As Statistics Netherlands wants to limit the number of IT applications to be developed and maintained, part of the SDMX project was to look closely at SDMX software already available. We found that quite a number of implementations are actually available with varying status and professionalism, both open source as well as proprietary [3]. We decided to look more closely at two of the most promising frameworks that were actively being developed at the time: the Istat SDMX framework [4] and the Eurostat SDMX framework [5] (also known as the Eurostat reference infrastructure). We studied the documentation, performed smoke tests, functionality tests and some limited performance tests. We also did some experiments with the Eurostat SDMX converter.

The outcome of this exercise was that both frameworks were usable. The Istat framework was easy to install and to use and had a very good web-service performance. The Eurostat framework offered more functionality, such as mapping assistance, and looked more promising overall. Our conclusion was that the Eurostat framework was better suited for our needs for building SDMX functionality in our dissemination infrastructure. In addition, the SDMX converter seemed usable as a robust tool for on-the-fly transformations.

However, tooling is not the only challenge in becoming SDMX capable. In fact the organisation of an SDMX based dissemination process, including management of international metadata like DSDs and the mapping of these 'external metadata' to 'internal metadata', seems at least as challenging. Certainly if we demand that the architectural approach matches the way our office - and the international dissemination process in particular - is organised. In other words, the technical and organisational approach should correspond to the business architecture. So before taking further decisions on this, we decided to write down the business architecture for international SDMX-based data communication processes in more detail. This work is still in progress.

4. A Dutch node in the Census Hub?

In the meantime the use of SDMX moved forward. The Census 2011 is one of the first statistical domains for which NSIs – including Statistics Netherlands - are required to deliver their results in SDMX [6]. In addition to this formal obligation, NSIs have been asked to take part in the Census Hub infrastructure [7]. In a way, the Census Hub can be seen as an improved version of the web services in the original SODI project mentioned before. The definition of web services in the SDMX specification has been improved and, compared to the situation in 2005, generic software is now available on the market. At this time, the implementation of the Census Hub is targeted at a single statistical domain only (which is why it is called Census Hub and not, say, Statistics Hub). However, [8] this infrastructure is expected to evolve into a generic solution to be used for other domains as well.

With respect to the Census Hub, in November 2010 Eurostat issued a first test DSD for one of the 60 hypercubes defined in the Census. Since Statistics Netherlands performs a register-based Census, we were able to compose a test dataset for this DSD. First of all, we used the SDMX converter to convert the test dataset into a valid SDMX file. It took some effort to transcode the data into the requested codes and to map the internal variables to variables in the DSD. However, it was certainly feasible. Secondly, we used the Eurostat framework to create a test web service on this cube with the same data. Although the Eurostat framework uses a different syntax to store mappings and transcodings, we were able to reuse them from the converter experiment. Of course it would be useful if SDMX tools used the same formats to store mappings and transcodings.

We concluded that is was technically possible to implement a Dutch node of the Census Hub. In fact, we discovered multiple scenarios to do so, all with their own pros and cons, but we could certainly choose one if applicable. It would, however, require resources to implement the solution in a production system. And, especially as the hub must be operational until at least 2025 [6], in addition to the one-off costs of setting up the solution, we would have to expect yearly costs for maintenance. All in all, we concluded that if the hub is to be used for one statistical domain only, so that Statistics Netherlands has to maintain multiple solutions for international data reporting, we do not have a valid business case to participate. However, if the hub were to develop into a general usable infrastructure for international communication of statistics, we do have a positive business case. At time of writing Statistics Netherlands has still to decide whether it will actually participate in the Census Hub.

5. Best practices from other domains

As part of modernising international data reporting, we also experimented with SDMX for reporting data in other statistical domains, such as statistics on fish, foreign trade, waste, and short-term statistics. As SDMX-ML evolved from Gesmes/CB and Gesmes/TS (also known as SDMX-EDI), the step was not that big. If there is a good description of the key family and code lists to be used, we were able to build the format. However, the advantage over SDMX-EDI is that the metadata are better organised. We therefore think that the introduction of SDMX-ML as the standard to be used as a universal format for the worldwide statistical community is a big step forwards. Nevertheless, simply transforming existing Gesmes based data flows into SDMX-ML alone is not enough.

As a result of our experiments, we got the feeling that, in order to make SDMX-based data transmission between (international) institutions and NSIs effective worldwide, we need to add some rules and best practices. Some of them might seem straightforward, but we still think we should mention them here:

- Before implementing a new SDMX-based data flow for a certain domain, try to re-use (or even refer to) what is already there for other domains.
- Use standard code lists if possible. There is always a good reason to invent your own, but from the perspective of coordination you have to have a very good reason to do so.
- Use a naming convention, such as found in [9], for code lists used in international transmissions. We have seen code lists where the same name (for example cl_area) refers to different lists, which can be confusing.
- It is good to work with human readable versions of DSDs called MIG (message implementation guide) or DTG (data transmission guide). However, we should agree on a standard format for these, too.

These are only a few simple observations from the point of view of practical use of SDMX for international data reporting. With the creation of the SDMX Statistical Working Group it might be possible to start formulating a more extensive set of best practices.

6. Some thoughts on the future of SDMX

All in all, we think that the introduction of SDMX is a step in the right direction. The use of an xml-based, easy to parse, internet-based language with extensive support for statistical metadata is clearly an advantage above sending excel sheets or csv files by email, referring to loosely coupled metadata specifications. The strength of SDMX is that is specifically targeted to the needs of official statistics with all the necessary metadata involved.

However, we also have to be honest about some of the downsides of SDMX. If you compare it with other web standards for handling multidimensional data (for example the Dataset Publishing Language (DSPL) [10]), it is clear that SDMX is rather complicated by design. This might explain the rather slow introduction. Also, the fixes that had to be made from version 2.0 to 2.1 [11], some of them on essential statistical concepts such as changing the overall structure of a key family considerably, indicate that earlier versions were not mature enough. But the positive view on this is that these fixes indicate that SDMX is emerging from a rather theoretical standard to a standard that is growing towards practical acceptance by statisticians.

So we feel that a lot has still to be done. First of all for SDMX to become a success, there should be a continuous effort to keep it simple and limit the standard to the exact needs of the statistics field, i.e., keep the focus on data and metadata exchange. If there are to be new versions of SDMX, they should be in the direction of an 'SDMX light' rather than adding new constructs.

Secondly, we feel it is essential to continue experiments with SDMX-based data communication between NSIs and international organisations and among international organisations on real datasets from multiple domains. The improvements in SDMX 2.1 which resulted from practical experiments show its usefulness.

And last but not least the standard should be used to solve problems in statistics that need to be solved anyway. An example is the international coordination and harmonisation of statistical metadata. Although SDMX theoretically facilitates coordination of metadata by means of a registry, its actual use for harmonising metadata across statistical domains and across international statistical organisations seems still very limited. This is an area for improvement that in our view could also positively influence its acceptance.

References

[1] *Automated Access to 100,000,000 Statistical Facts via StatLine4 Web Services*, Invited paper, Joint ECE/Eurostat/OECD Meeting on the Management of Statistical Information Systems (MSIS), April 2005

[2] *39th Session of the UN Statistical Commission,* New York, Feb. 2008

[3] *SDMX Tool Repository*, http://www.sdmxtools.org

[4] *Istat SDMX framework project*, http://bms.istat.it/sodidownload/Download.aspx

[5] *Estat reference infrastructure,* Eurostat

[6] *Commission Regulation (Eu) No 1151/2010* of 8 December 2010

[7] *5th Meeting of the European Statistical System Committee (ESSC) 63rd EEA Conference*, Luxembourg, 20 May 2010

[8] *Meeting of the Census Hub IT working group*, Feb. 2011

[9] EDAMIS Naming convention for data senders use, Eurostat unit B5, v. 5.01, 4 January 2010, http://circa.europa.eu/Public/irc/dsis/edamis/library?l=/reference_documents/edamis/dataset_convention/_EN_1.0_&a=d

[10] *DSPL: Dataset Publishing Language*, http://code.google.com/apis/publicdata

[11] SDMX *2.1 Summary of Major Changes and New Functionalities*, 1 December 2010, SDMX consortium