Distr.
GENERAL

WP.11
30 April 2011

ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

Meeting on the Management of Statistical Information Systems (MSIS 2011)
(Luxembourg, 23-25 May 2011)

Topic (ii): From local to corporate perspective (industrialization and standardization)

Improving interoperability in Statistics – The impact of SDMX Some considerations

Invited Paper

Prepared by Rune Gløersen, Statistics Norway[1]

I. Introduction

1. National Statistical Institutes (NSIs) are currently reshaping their business processes in order to meet the increasing demand for high quality statistics. At the same time NSIs are facing budget cuts, intensified competition, consequences of globalisation and request for reduced response burden. NSIs are to some extent squeezed between non-compatible demands. There is no single measure to be taken in order to meet such a variety of challenges. Different initiatives need to be aligned. The challenges are common to the statistics community, indicating that the challenges should be met by the community, not separately by each individual organisation. In order to act as a community, there is a strong need for alignment and increased interoperability between the members, and between the community and the outside world. The increased adoption of SDMX both at the level of NSIs and among the international consumers and producers of statistics is one of the important efforts to improve the business of the statistics community.

2. The ambition to increase interoperability and to align business processes is evident within the statistics community. One example is the high-level decision to use SDMX as the format and mechanism for exchanging statistical data and metadata between members of the community. Furthermore, and more ambitious is the example of the Eurostat Census Hub project, where the end user is supposed to obtain census tables by defining queries that crawl the individual NSI websites for data, in order to compose one consistent resulting table. Taking this one step further, the same functionality will be generally demanded for putting together cross-national tables within or across the statistical domains. In addition, projects have been established to describe common architectures and high level models, as a foundation for further efforts to standardise business

[1] The views expressed in this paper are those of the author, and should not be attributed to Statistics Norway

processes and process means. However, standardisation is not a goal in itself, and we should be precautious in order to avoid counter productive solutions or spending efforts competing with the commercial industry.

3. NSIs collect data and produce statistics in a complex environment, interacting with a broad range of users representing a variety of needs or obligations. We are currently in a change from collecting and disseminating data in strictly defined data flows, to collect and integrate data from structured databases/registers, electronic traces or non-structured information sources among the vast volume of data created on the internet. The production of statistics is to a less extent than before controlled by the statistical institutes. Our output data are rapidly utilized and composed into new information widely available. The implication of this evolvement is that we are dependent on the adoption of standards in adjacent environments, and should investigate the coherence with our statistics community standards. The Linked Open Data initiative is one example in this area, the growth of Google being another. We should consider our potential influence in the development of such related standards.

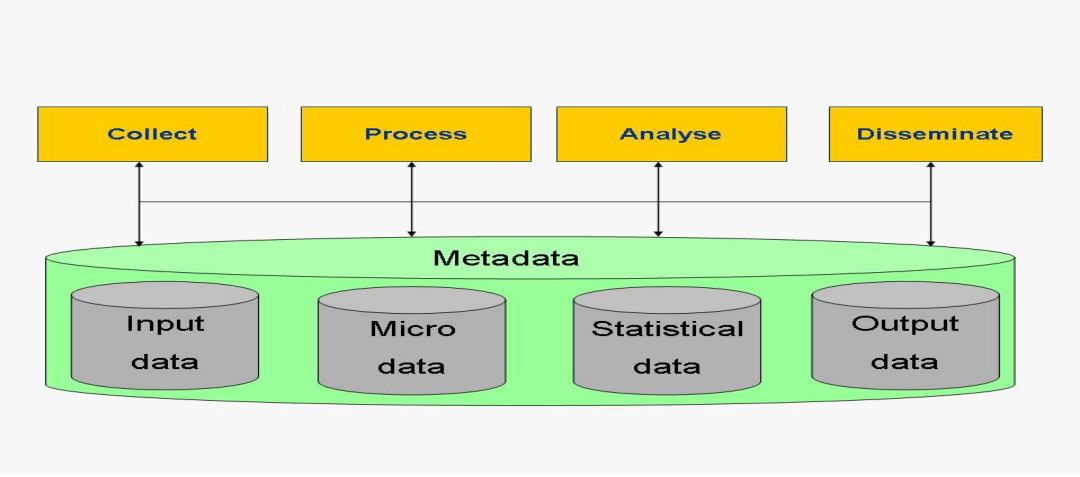
4. Starting from being a format for exchanging aggregated data and associated metadata based on well defined agreements between identified partners, SDMX is currently being investigated for use in a lot of areas throughout the statistics production, and is also expected to play a role when utilizing commercial tools used in the production process, for dissemination, visualisation etc. The objective of this paper is to contribute to the discussions in this area.

II. Streamline production and standardize data flows

5. The advantage of standards linked to the statistical business processes is discussed in the following. The reference model is the Generic Statistical Business Process Model (GSBPM). In coherence with the statistical business processes, data are transformed through a number of stages from raw data to published statistical information. There is ongoing work in several NSIs to define and implement common data stores for archiving data at steady states of the data life cycle (i.e. in Statistics Sweden, Statistics Netherland, Statistics Norway and others). However, the number of steady states is not well defined or commonly adopted. In the figure below, 4 states are described, linked to the 4 core business processes. The data flows between the processes, and especially the reuse of data from the different stages shows a considerable difference with respect to the nature of the data and the objective of the data exchange for further usage.

6. The description in figure 1 is simplified, and intends not to cover in detail every aspect of the statistics production. The aim is to give a view of the diversity of the data handled by NSIs, and the broad range of primary and secondary usage.

Fig. 1 Characteristic states of data linked to the core business processes



7. Streamlining the statistics production implies a need for standardising the data flows between the different business processes, broken down into defined structures of the data and metadata required as input and expected as output from the different (sub) processes.

A. Collect

8. Data collection is carried out in interaction with various types of respondents, using different technologies and channels. In addition, data is collected as reports or copies from administrative registers or business systems, or by capturing information available on the Internet. The volume of data collected through the various channels varies between countries.

9. In order to keep the response burden low while keeping the data quality high, data should be collected in a way that requires minimum efforts by the respondents. This means that we need to adjust our data collection instruments according to the preferences of the respondents, and to look for possibilities to capture data that already exists, either as reports from existing sources, or by hooking up to existing flows of information.

10. A number of standards are already in use, and have been investigated for the purpose of enhancing data collection by NSIs. Some examples are

- XForms, an XML specification to define web forms. XForms has the status of being a W3C Recommendation, and are supported by various tools (http://www.w3.org/MarkUp/Forms).
- XBRL, an XML specification for business reporting. Widely used in some countries, and often referred to as a standard with a potential to automate data collection some areas relevant for NSIs (http://www.xbrl.org)

11. The pressure to reduce the response burden is likely to continue. Standards exclusively used by the statistics community for data collection will probably not meet market or end user expectations. Data collection at national level is also subject to national co-ordinating initiatives, like e-Government. The consequence is that when commercial technologies and standards reach momentum among businesses or citizens, the statistics community must map their requirements by bridging public or industry standards with statistics community standards.

B. Process

12. The main objective of this process is to transform Input data into clean and quality assured Micro data. Several sub processes comprising data integration, coding, editing, imputation, estimation etc. contribute to the final resulting micro data.

13. The NSI Micro Data Archive is subject to extensive secondary usage, especially for research purposes. The business of keeping a well documented micro data archive is to a large extent common between NSIs and National Data Archives. The micro data are normally kept as the observed values, or restructured into micro data targeted for further utilization, i.e. integration, analyses.

14. There are a number of standards available to structure, document and exchange these archived micro data. Some examples are

- ISO 11179, a standard to establish *Metadata registries (MDR)* which addresses the semantics of data, the representation of data, and the registration of the descriptions of that data (http://metadata-stds.org/11179).
- Data Documentation Initiative (DDI), an effort to create an international standard for describing data from the social, behavioural, and economic sciences in XML. (http://www.ddialliance.org)

15. Some members of the statistics community launched in 2004 the Neuchâtel Terminology Model for classifications (version 2.1) and variables (version 1.0), as an attempt to create a standard in this area (see Metis

Wiki at http://www1.unece.org/stat/platform). Additionally, SDMX is also used for exchanging micro data, i.e. in the EGR (European Group Register) domain.

16. Secondary use of micro data, especially for research is typically supported by standards the have the intention to structure and describe the metadata to provide for the best possible understanding of the semantics and quality of data. Since distribution of statistical micro data is heavily restricted, access to micro data for research is often controlled by the owner of the archive, which means that a standard for exchange of data is less relevant, while a standard for describing the data is more important.

C. Analyse

17. The main activities in this process are the aggregation into the numerical basis for the domain specific statistics, and the calculation of the derived metrics like relative figures, indexes etc. Furthermore, to assure quality, interpret and explain the statistics.

18. The data archive is labelled Statistical data. The reason for this is that the archived data at this stage usually is a combination of micro data and aggregated data, constituting a statistical domain data warehouse designed according to the characteristics of the domain, user needs taken into account.

19. Aggregated statistical data is subject to disclosure control etc. at this stage, before being left over to the output data base for public dissemination. The data warehouses at this stage contain data targeted for specific use and analyses. Normally, exchange of data between partners, like reporting data to international organisations based upon specific requirements will be extracted from these data archives, by specific routines according to the agreed specifications. This type of data exchange represents the first and still most common use of SDMX.

20. The reason for using the Statistical data archive as the source to build the aggregated data sets could be due to special requirements, like code lists that defers from the code lists used for the common published statistics.

21. SDMX is a standard well suited for this type of data exchange. The transmissions are stable, repeated over time, and the richness of the exchange format is a good foundation for automating not only the transfer process, but also the further processing of the information. Alternative standards are far simpler and usually just format specific without the same support for providing associated, content oriented metadata.

D. Disseminate

22. The final stage of the statistics production is dissemination of controlled and publicly available data. Even though dissemination, as described above, also comprises dissemination of micro data or aggregated data subject to licensed use, this paragraph covers only the dissemination of public macro data.

23. There are a number of existing standards and initiatives for the exchange of statistical output data. Some examples:

- SDMX is one obvious standard for this purpose. However, for general purpose output across all domains the implementation method of SDMX seems to be varying.
- Standards defined to enable the vision of the Semantic Web, like Resource Description Framework (RFD) and ontology languages like OWL.
- The newly available and still evolving Dataset Publishing Language (DSPL) from Google.

In addition, we have statistics community standards, like for instance the PC-Axis (PX) files.

24. Data are not only disseminated as stand alone aggregated data, but are fed into commercial tools like spreadsheets, OLAP tools for further analyses etc. The statistical data are also integrated with other types of data, for instance using visualisation tools to create statistical information on maps etc. Hence, in this area we

are also facing the need for integration of standards for statistical data and standards designed for different purposes, like geographical data (for example GML). This underscores the need and value of commonly adopted standards to describe and exchange statistical data.

III. Some observations

25. There are a number of standards to describe, format and exchange statistical data for different purposes available. The standards are partly overlapping and partly complimentary. Nevertheless, there is a clear objective to encourage the NSIs to use SDMX in most cases. But the benefits of this advice are not completely clear.

A. Lessons learned

26. The first standards created for the exchange of aggregated statistical data were released more than 20 years ago, by the introduction of the GESMES Edifact message. The basic concepts were brought forward into the development of the SDMX standard sponsored by seven international statistics agencies. However, the adoption of the standards has been rather slow within the statistics community, and even slower outside. The intention to create generic standards is always quite ambitious. The complexity will usually increase; leading to criticism like that the standard is too complex for a specific task, but still lacks functionality to become completely generic and useful in any relevant area. This will lead to continuous improvements in detail, contributing to increased complexity and vast amount of documentation.

27. The ambitions in creating a standard for generic data exchange should also be assessed with respect of the suitability of the collaborating systems. Specialised systems, for instance registers handling micro data are normally not advanced metadata driven. The systems handle complex internally coded business rules, while the exchanged data flows often are rather simple. Using generic mechanisms for the exchange of data and metadata could initiate inappropriate complexity in the routines needed to compose and decompose the transferred information. On the other hand, a generic approach to exchange aggregated data should in most cases be highly appropriate, because aggregated data (tables or cubes) by nature are objects on a sufficient higher level of abstraction.

28. Introducing standards for exchanging data implies the need for interfaces. Data at its origin is transformed through an interface, transferred and submitted to the receiving interface. Well defined interfaces should lead to an opportunity to optimize internal work processes and the process of exchange. If these interfaces end up being wrapped in some kind of black-box software, the flexibility to adapt to the internal work processes on either side of the cooperating environments could end up being reduced. In such cases, the data exchange could end up as standardised, but from a process point of view, disconnected islands.

29. If we look at the successes of standardisation, in most cases success is related to complexity. The European Central Bank had a success using Gesmes/TS some ten years ago. Gesmes/TS was a simplification of the original Gesmes message targeted for the purpose of the exchange of time series. The various standards that are rapidly evolving on the internet are normally rather simple by nature, leaving wide opportunities to create targeted solutions with high business value. Concerning technological standards, we have very often experienced that it is not the most sophisticated standard that penetrates the market, but the simplest one, like Ethernet, TCP/IP, SMTP etc. This can also be referenced by the 80/20 paradigm; the tendency to spend 80 per cent of the total resources, solving the last 20 per cent of nitty-gritty functionality perhaps needed.

30. The developments of standards are driven by specialists, while standardisation must be business driven. This means that any standardisation efforts should be anchored sufficiently to the businesses involved, and become subject to thoroughly analysed business cases. In this respect, we can say that most of the barriers to success have not been caused by the standard or technology itself.

B. Increase interoperability

31. A successful adoption of standards within the statistics community is one of the key drivers for enhanced interoperability. However, interoperability will increase over time and mature in stages. The need to share data is a typical starting point for increased cooperation. This has increased the efforts to use standards and share the tools used within the relevant work processes, and intensified the cooperation needed to develop them. Nevertheless, we are still at the beginning of the steps leading towards more comprehensive interoperability within the statistics community.

34. Increased interoperability is basically based on three pillars:

- Organisational aspects, like similarities in the legal basis of the businesses, globalisation, comparability, harmonisation of products, willingness at managerial level to collaborate based upon common objectives or strategies etc.
- Semantical aspects, like commonly used content oriented standard definitions and classifications, the need to harmonise information, outputs etc.
- Technical issues, like common use or development of tools, technical standards and technologies

35. Recently, a number of initiatives have been launched in order to improve the harmonisation of products, methods and production means in statistics. It has been acknowledged that it is becoming too expensive for each and every NSI to individually change their tailored production systems to meet emerging technologies, standards and user expectations. Accordingly, the approach to harmonisation is changing from bottom-up and domain specific actions, to better coordinated, top-down management driven initiatives.

36. If we look at the three pillars of interoperability from the view of the statistics community, it will briefly cover

- The way we run our business, i.e. the business processes
- What we are producing, i.e. statistical data and analyses
- The methods and tools that support the business

37- The UN/ECE METIS Group released the GSBPM a couple of years ago. The model has been widely adopted as a reference model that sufficiently describes the business processes of the Statistical Institutes. The model is kept on a rather high-level, and does not have the quality to become completely prescriptive as a standard. However, this high-level model has made significant contributions to the work on further harmonising statistics production. By using the model, we can now label specific tools according to a more precise and commonly adopted description of the processes supported. This adds a unique possibility to increase the sharing of software. However, it has become equally clear that there is a need for a similar, high-level model of statistical data, which should take the same role as a reference model.

38. If it is possible to require that any flow of data should be consistent with a conceptual model, the model itself could act as a bridge not only between different flows, but hopefully also between relevant standards. Achieving this would imply a greater flexibility in designing more targeted solutions on a detailed level. A top-down approach to data and metadata standardisation should start with an agreement on such a conceptual level, to achieve the commitment and governance needed from the participating organisations

39. The shift to acknowledge the value of agreed, common models as the basis for enhanced cooperation is one of the indications of enhanced maturity of interoperability. The phenomenon has been observed as an evolutionary path also within e-Government interoperability. The four stages describing this evolution are[2]

- *Stage 1 - Aligning work processes*. Efficient operation requires integrated activities, schemas and data exchange based upon detailed specifications among different information systems.
- *Stage 2 - Knowledge sharing*. Agencies put effort into defining best practises, specification of metadata, methods and technical standards for information systems and data exchange.

[2] Hans Solli-Sæther, Analytical Framework for e-Government Interoperability, www.echallenges.org and also www.semicolon.no

- *Stage 3 - Joint value creation*. Common information models and service catalogues are developed to facilitate joint developments of services to the benefits of the end-users.
- *Stage 4 - Strategic alignment*. Common strategic positioning among agencies, and adaption of laws and regulations.

The statistics community is currently moving to stage 2, and starting to scratch at stage 3.

40. The decision to use SDMX for (all) data communication that was endorsed by the 39th Session of the UN Statistical Commission was in a way a top level decision over a bottom up approach, leaving implementations to be executed on data flows on domain levels. However, SDMX also comprises two important parts which should have been pushed more clearly as part of the decision. This is the underlying Information Model, and the overall Common Metadata Vocabulary (CMV).

41. During the comparison of the DDI and the SDMX standards, it is the overlaps and differences at the conceptual level of the models that is most interesting. From a statistics community point of view, a model comprising both, but still being kept at a minimum level of details could act as a common information model for the statistical data processed and disseminated by Statistical Institutes. This is basically the idea behind the development work started among some NSIs led by the Australian Bureau of Statistics with the aim to develop a Generic Statistical Information Model (GSIM). Accordingly, the CMV could be elaborated to become a common vocabulary for the processes, metadata and methods within the statistics community. This is another contribution to a high-level harmonisation, and would for instance ease the work that a lot of different Statistical Institutes put into developing their own thesauruses.

42. The efforts made on agreeing on harmonised high-level models are encouraging standardisation in a number of areas. Standardisation projects have the objective to describe how to reduce or eliminate the unnecessary variations in work processes, metadata definitions and statistical methods. Even though it is quite common to interpret standardisation as activities that reduce flexibility and remove creativity, standardisation should be regarded on the contrary; as an enabler to release resources and strengthen improvements by moving resources from manual or inefficient work, to create new products or services. Improved harmonisation within the statistics community will also increase the market and business value for commercial developments. This type of standardisation efforts should lead to an industrialisation of statistics, a term pointing forward in modernising our community.

IV. Conclusions

43. SDMX has been one of the important drivers to strengthen the efforts to harmonise the production and dissemination of statistics. However, the development and implementation of the standards have been driven mainly from the point of view of the seven sponsors among the international statistical organisations. While the standard fits very well for the exchange of data between statistical agencies, it is debatable whether it is suited for all processes and flows of data, especially at the level of NSIs. In addition, the standard overlaps with other standards relevant for statistics, like the DDI, and also when compared to existing and emerging standards on the internet.

44. SDMX comprises elements of high value to improve the overall harmonisation of production and dissemination of statistics. These elements should be elaborated to become high-level conceptual models, accepted as common reference models. This should leave an opportunity for more flexible, and in some cases less complex use of relevant standards. Furthermore, it will make Statistical Institutes less vulnerable to change in standards, and open up for an easier, but still controllable adoption of evolving commercial standards.

45. Sufficient technical interoperability is the underlying objective of any standardisation effort, in order to automate processes and implement efficient IT services. However, technical developments must be driven by the business, and implementation of ready made tools should be accordingly assessed. To ensure this, any standardisation projects should be based upon well defined and approved business cases.