Information management as tool for standardization in statistics

Barteld Braaksma[1,2], and *Kees Zeelenberg*[1]
[1] Statistics Netherlands, The Hague, Netherlands
[2] Corresponding author; email: b.braaksma@cbs.nl

Abstract

We describe the role of information management in redesign programs: why it is needed, how it is at present being developed through international cooperation in the Generic Statistical Information Model (GSIM) project, and we give some examples how information management has been used in practice at Statistics Netherlands. We conclude that GSIM is necessary for more extensive usage of big data sources, for interoperability of statistical tools and for sharing tools between NSIs.

Keywords: information management, standardization, business process redesign, official statistics

1. Introduction

Many National Statistical Institutes (NSIs) have initiated redesign programs of their production processes for official statistics. There have been several external and internal causes and drivers for these modernization programs. Externally NSIs have been faced with budget cuts, in the last years even severe ones, as well as demands from users for greater transparency, reducing the burden on respondents, and greater flexibility and new and more timely statistics. Internally NSIs need to get more grip on the information systems and processes that underlie the statistical production processes, and must increase labor productivity.

In recent years, new drivers for modernization have come from the political demand for more timely short-term indicators and for higher quality of macro-economic statistics as well as the advent of *big data*.

From the experiences with those redesign programs, it has become evident that these programs must be structured and coordinated in various ways in order to be successful and not defeat their own purposes. In turn, this has led to various standardization initiatives, both within NSIs and in joint cooperation between NSIs, relating to business architecture, methods, tools, project management and information management.

In the next sections, we first discuss briefly the backgrounds of the modernization trends in NSIs and the various standardization initiatives. We use Statistics Netherlands (SN) as an example, but also refer to international initiatives at the level of the European Union (EU) and the United Nations (UN). The main part of the paper deals with the role of information management in redesign programs: why it is needed, how it is at present being developed through international cooperation in the Generic Statistical Information Model (GSIM) project, and we give some examples how it has been used in practice at Statistics Netherlands.

2. Modernization in official statistics

Modernization programs

As described in the introduction, NSIs have come under increasing pressure to modernize their statistical production process. Many NSIs have introduced modernization programs, aimed at modernizing the main processes, such as business statistics, national accounts, and household statistics. The objectives of these programs have often been: 1. improvement of quality, in particular with respect to coherence, flexibility, consistency and transparency and reproducibility; 2. more use of administrative data in order to reduce respondent burden; 3. higher efficiency; 4. reduction of the number of IT applications.

Examples are the *ABS 2017* program (Borowik, Henden, and Fraser, 2012; Hadley, 2013) at the Australian Bureau of Statistics (ABS) and the *Counting on Statistics* program (Ypma and Zeelenberg, 2007) at Statistics Netherlands, and there are other examples at many other NSIs. The experience with these programs has shown that an integrated approach is necessary, both in design and development of the statistical processes and systems and in the actual operation of the processes. In the redesign of the statistical processes, this integrated approach has led to the development of enterprise architectures for NSIs, which usually consists of four parts: the business- and information architecture parts and the applications and infrastructure architecture parts. The business architecture describes the business goals including desired outputs and the strategies to achieve these goals and makes them sufficiently specific. The other parts of the architecture can basically be considered as a framework for IT solutions supporting the business goals and strategies, which provides guidance to the IT.

At the international level, we see similar pressures for modernization, in particular the demand for higher quality of macro-economic statistics, such as the government deficit and debt. In the European Union this has led to various quality and standardization initiatives, such as the Code of Practice for European Statistics and the Sponsorship on Standardization (Braaksma *et al.*, 2013).

Statistical value chain

In the business architecture developed at Statistics Netherlands (Struijs *et al.*, 2013), the statistical production process is modeled as a *value chain* of subprocesses, such as, at the highest level, data collection, editing, integration, and dissemination; of course these can and should be subdivided in more detail. The output of a subprocess is the input of the next subprocess; this can be thought of as information objects flowing from one subprocess to another. To keep the system manageable, not all steps in the value chain need to be considered equally important. Those that are the most important are called *steady states*; to be more precise they are points in the value chain where a complete set of data with accompanying metadata and of a predefined quality level, is maintained; in other words: value is added at a steady state. At least 6 steady states are identified: the *pre-input base* for raw sources, the *inputbase* for source data as they enter statistical processing (editing, imputation, integration), the *microbase* for statistical micro-data after statistical processing, the *statbase* for aggregated statistical data, the *output base* for data that may be disseminated (i.e. after statistical disclosure control has been applied), and the *post-outputbase* for information that has actually been disseminated, which includes the published data but also press releases, analyses and other context information.

An important aspect of the statistical value chain is *management of the chain*: the production process as a whole needs to be planned and controlled according to well-defined quality aspects for the steady states. The steady-state concept is thus closely related to that of *quality gates* (ABS, 2010).

An example at Statistics Netherlands is the chain of economic statistics, which ranges from the business register to the national accounts. In recent years this chain has been redesigned completely according to the business architecture of Statistics Netherlands. Of course, the interaction between architects, methodologists and statisticians has also led to more clarification on the possibilities and limitations of such an approach. For example, it has appeared that the actual chain looks more like a network, where the main line has several branches. Further, the importance of steady states and of chain management clearly emerged. Also it took some years and several experiments before we could design the final organization that best fits the chain.

3. Information objects and flows in the statistical value chain

Information objects and flows

What NSIs really do is transforming and transferring information, i.e. the statistical value chain can be regarded as a flow of information objects. Roughly speaking, each statistical production process starts by taking some input data, processes it and delivers a set of output data. This holds on any level of granularity. E.g. when considering a national statistical institute as a whole, its input consists of all survey data it collects, together with more or less raw data obtained from other data collectors like the Tax Office or Internet sites; these data are then processed in order to compile tables for publication on its website or for use in press releases. For an international organization like Eurostat, the OECD or the UN, the input data may consist of finalized aggregates from several countries which are then combined into meaningful statistics describing for example inflation in the Eurozone. At a lower level, e.g. an editing or imputation process, input is unedited data and output is improved data after editing or imputation. The common element in these examples is that a set of data is transformed into a different set of data that has a higher statistical value. Either the data itself is transformed or the metadata that describes its characteristics. In a broad sense that even applies to seemingly trivial operations like copying and pasting a dataset between IT tools.

Examples of information objects and flows

There are different ways and levels to describe information and associated metadata that is passed from (sub)process to (sub)process. We give two examples from recent redesign projects at Statistics Netherlands.

The *HEcS program* was a large-scale redesign program of the whole chain of economic statistics that ran from 2006 until 2011 (Braaksma and Buiten, 2012). When the program started, a high-level architecture was defined that basically identified steady states in the process chain, or rather process network. For these steady states it was described what kind of data they contain, including some key characteristics like quality attributes. Since its inception this steady state architecture has served as a backbone for the whole program. For example, the scope of development projects under the program has been defined in terms of steady states; processes, methods and IT-systems have been developed to transform data between steady states; and even the organizational structure of the HEcS production units has been adapted to reflect steady states.

A second example is the *MUST project* that aimed at redesigning the Environmental Expenditure Business Statistics. This redesign was long overdue: first attempts already started ten years ago but for several reasons never materialized. When the project started in 2011, a very systematic business analysis (Renssen, 2012) was carried out that identified the required outputs and available inputs and that divided the production process into several logical steps in order to define the *flow* of the whole process. Only after these steps were defined, the *know*, i.e. the methods used in each step, were selected. When both processes and methods were sufficiently clear, an IT information analysis was done to decide how to implement the design in an efficient way. It turned out that on a technical level, in some cases it was useful to combine several logical steps in a single IT-component. The system has been implemented in SPSS, although R was also considered, and is now in production. Key development documents in this approach are the BAD (Business Architecture Document), MAD (Methods Advisory Document) and SAD (Software Architecture Document); see Hofman (2011).

GSIM as a language for information flows

Since the production of official statistics entails flows of statistical information, it is useful to have a common language to describe what happens to these flows of information objects, especially when several actors are involved; in technical terms, such a language is called an *information model*. This applies in particular when actors belong to different organizations, but even when they belong to different units within the same organization it is useful. And above all, when (re)designing a statistical production process, it needs to be clear to all parties involved what is happening and what they can expect. When we started with the HEcS and MUST projects given as examples above, a common language sufficiently rich to describe all needed information objects in a standard and formalized way, was still lacking.

The *Generic Statistical Information Model* (GSIM) is such a common language. It enables communication by offering a standardized set of key statistical information objects that apply throughout the statistical process. It thus supports standardization in “thought” and thereby makes collaboration easier. In the longer run, using the same language assists and may even force the standardization of processes. For this it is necessary that this language, GSIM, is established as a common standard; the ESS Sponsorship on Standardisation has described the process, rules and procedures that lead to establishing a standard, see Braaksma et al. (2013).

An example how GSIM assists standardization, is the current CSPA (Common Statistical Production Architecture) initiative, which tries to define what a statistical Plug & Play environment would look like. In this environment, a statistical process can be built ‘LEGO-wise’ from several components that are made to fit together. It is obvious that this implies that the information that passes through such a component must be well described, including associated metadata and processing parameters.

A somewhat simpler example is SDMX (Statistical Data and Metadata eXchange). This was specifically designed to facilitate data exchange between statistical organizations, typically national statistics producers on one side and international organizations on the other side. Based on the SDMX specification, several software solutions have already been developed. Differently from its precursor GESMES (Generic Statistical Message), SDMX has been adopted by a number of international organizations who have founded the SDMX Sponsors Group, as their preferred standard. In the European context, the use of SDMX is even being incorporated in legislation. This indicates a major prerequisite for a standard: enough relevant parties must support it and actually use it.

A similar example is DDI (Data Documentation Initiative), which is an effort to create an international standard for describing data from the social, behavioral, and economic sciences. It has been developed and supported by a broader community than official statistics, including e.g. universities, research centers and data archives.

Since both SDMX and DDI are already used in the statistics community, one of the challenges for GSIM was to provide a model that is compatible with both DDI and SDMX. Of course, there are numerous other data description standards that are in broad use like Google’s DSPL (Dataset Publishing Language). It is impossible for GSIM to incorporate them all, which means that we still need interfaces and transformation mechanisms.

The difference between GSIM and the examples mentioned above is that GSIM tries to be model all the information objects used in statistical production processes. For example, DDI does not detail the paradata, the information about processing and production of statistical data. SDMX focusses on the description and exchange of aggregate data, whereas GSIM is more general.

On the other hand, DDI and SDMX are in some respects more detailed than GSIM, and so can be used to implement parts of GSIM. For example, DDI has a very detailed description of the survey instrument and of data collection, whereas GSIM leaves this more abstract. Similarly, SDMX might be used to provide more detail to the part of GSIM that deals with exchange of statistical data.

Because GSIM, DDI and SDMX are partly complimentary, cooperation and integration is useful, and indeed work is under way to achieve this.

GSIM 1.0 only appeared at the end of 2012, after HEcS and MUST were finished. If GSIM had been available before, we might have used the GSIM language to describe the HEcS steady state architecture, and the *flow* in the BAD of the MUST project. Of course, the *know* is beyond the scope of GSIM and requires other tools like the Statistics Netherlands’ Methods Series (CBS, 2013), which describes the validated methodology available for official statistics. When people are trained in the GSIM language, that helps to translate between different disciplines. Nowadays, still a lot of communication (including check/double check) is needed to make sure that for example an IT architect understands a business problem the right way and can come up with the right solution. On the other hand, experiences at SN with redesign projects have helped to shape GSIM. Key people who were involved in e.g. HEcS and MUST have also been involved in GSIM development activities. Their experiences have influenced their contributions to GSIM, in particular the part dealing with statistical production.

Implementation of GSIM

An open question remains how to implement GSIM. A first step is to start using the language of GSIM by teaching people involved in redesigns like business analysts, IT developers, methodologists and project managers its concepts, and making sure they reach a common understanding. This will help the design and redesign of processes. A further step is to start using, and contributing to, standard 'Plug & Play' style software components that will hopefully be offered in the future under the CSPA.

There is no business case for replacing old legacy systems just for the sake of complying with GSIM; in such cases, compliance with GSIM may be realized only when for other reasons a statistical process will be redesigned. Nevertheless there ìs a business case for developing from now on new systems and processes 'future-proof' in the sense that their inputs and outputs, including metadata and process parameters, are clearly specified and compliant with the GSIM language. In particular in cases where interfaces with other GSIM-compliant systems are expected this makes sense.

An area where it definitely makes sense to start using GSIM from scratch is in the incorporation of new technologies. Legacy systems are usually developed in a national context and taking national needs and opportunities into account. The main challenges of new technologies do not respect any national border: think of using mobile devices like smartphones and tablets for data collection or data dissemination. The interfaces need to be localized in different languages, e.g. as in the UNECE guidelines. But apart from that, the way these devices can be applied is universal. And in particular in a time of diminishing budgets it is very attractive to develop a joint approach.

The same applies to the oncoming Data Deluge. When we start to use Big Data sources like traffic-loop data, call-detail records from mobile-phone usage or Twitter messages for statistics production (see e.g. Daas et al, 2012, Daas et al., 2013, and the UN Global Pulse initiative), one of the main problems will be to extract meaningful statistical information from data sources collected for totally different goals. Big Data present NSIs with the need for more intensive cooperation with data providers, data collectors and data intermediaries. For example, it is often not possible and even if it is, not efficient, for an NSI to transfer all mobile-phone data to its data center. So we have to devise new ways to analyze these data, e.g. by drafting queries, that are carried out by a data intermediary on databases of mobile-phone companies. To make such cooperation work, we need a common language that specifies how information is stored and transferred. In other words GSIM is an essential prerequisite for cooperation. And since the whole statistical community worldwide faces the same challenges at the same time, it makes sense to start talking about them in a standardized way. So GSIM needs a 'Big Data' module. However, if this kind of use for GSIM does not materialize and GSIM is not suitable for elaborating the use of new technologies and new data sources, ad-hoc solutions for communication and cooperation will be searched for, with all the negative consequences, such as patch work and incompatibilities.

4. Conclusion

NSIs are faced with the challenges of standardization and integration of their statistical production processes, and increasingly with the challenge of international standardization and integration across NSIs. To meet these challenges, we must standardize the languages that we use to describe our statistical processes. The structure of these processes is described by the business and information architecture, for which we use GSBPM. Since statistical information flows are an essential element of statistical processes, we must also provide a description of these flows, for which GSIM has been developed.

We do not see a business case for replacing old legacy systems just for the sake of complying with GSIM; in such cases, compliance with GSIM may be realized only when for other reasons a statistical process will be redesigned. However, GSIM is a prerequisite for plug & play IT tools, international cooperation and the use of big data in statistics.

Looking ahead, we see 4 major developments:

1. Further elaboration of GSIM, in particular adding elements from DDI and SDMX, and possibly integration of these three information models.
2. Application of GSIM in the redesign of statistical processes within NSIs, and in the design of new processes, in particular those using big data sources.
3. Use of GSIM / CSPA in the development of plug & play software modules.
4. Use of GSIM to facilitate sharing tools across NSIs.

References

ABS (Australian Bureau of Statistics), 2010, *Quality Management of Statistical Processes Using Quality Gates*. Information Paper 1540.0, Australian Bureau of Statistics, Canberra.

Borowik, Jenine; Merilyn Henden; Bruce Fraser, 2012, *Riding the big data wave to streamline acquiring and managing data*. Paper presented at the Meeting on the Management of Statistical Information Systems (MSIS 2012) (Washington, DC, 21-23 May 2012).

Braaksma, Barteld, 2009, *Redesigning a Statistical Institute*. Paper presented at the Conference on Modernisation of Statistics Production, Stockholm, 2-4 November 2009.

Braaksma, Barteld; Gert Buiten, 2012, *Redesign of the Chain of Economic Statistics in the Netherlands*. Paper presented at the Fourth International Conference on Establishment Surveys (ICES-IV), Montréal, 11-14 June 2012.

Braaksma, Barteld; Cecilia Colasanti; Piero Demetrio Falorsi; Wim Kloek; Miguelangel Martinez Vidal; Jean-Marc Museux; Katalin Szep, 2013, *Standardisation in the European Statistical System*. Paper presented at the NTTS Conference, Brussels, 5-7 March 2013.

CBS (Statistics Netherlands), 2013, *Methods Series*.

Daas, Piet J.H.; Marco J. Puts; Bart Buelens; Paul A.M. van den Hurk, 2013, *Big Data and Official Statistics*. Paper presented at the NTTS Conference, Brussels, 5-7 March 2013.

Daas, Piet; Marko Roos; Mark van de Ven; Joyce Neroni, 2012, *Twitter as a potential data source for statistics*. Discussion Paper 2012-21, Statistics Netherlands, Heerlen.

Hadley, Patrick, 2013, *Building Organisational Capability for Plug and Play*. Paper presented at the Meeting on the Management of Statistical Information Systems (MSIS 2013) (Paris, 23-25 April 2013).

Hofman, Frank, 2011, *Redesign approach at Statistics Netherlands*. Paper presented at the 2011 International Methodology Symposium, Ottawa, 1-4 November 2011.

Renssen, Robbert, 2012, *Process Model of the Environmental Expenditure Statistics*. Paper presented at the Second GSIM Sprint, Daejeon, 16-27 April 2012.

Renssen Robbert; Mattijn Morren; Astrea Camstra; Tjalling Gelsema, 2010, *Standard Processes*. Statistics Netherlands Discussion Paper 2010-13.

Struijs, Peter; Astrea Camstra; Robbert Renssen; Barteld Braaksma, 2013, Redesign of statistics production within an architectural framework: the Dutch experience. *Journal of Official Statistics* 49(1), pp. 49-71.

Ypma, Winfried; Kees Zeelenberg, 2007, *Counting on Statistics; Statistics Netherlands' Modernization Program*. Invited paper presented at the Conference of European Statisticians, Geneva, 11-13 June 2007.