VERSITA

Journal of Official Statistics, Vol. 29, No. 1, 2013, pp. 147–155, DOI: 10.2478/jos-2013-0008

Discussion

Daniel Defays[1] *and Jean-Marc Museux*[1]

1. Preamble

This discussion draws mainly on four contributions in this special issue (those of Statistics New Zealand, NASS, RTI International and Statistics Netherlands) that report on experiences in integrating statistical production systems with a direct focus on industrialisation and integration-related issues. The framework proposed by Eltinge et al. for the integration of architecture and methodology is also taken on board. The article on the Real-Time Online Analytic System (Westat/NCHS) is marginally integrated into the discussion on data access.

2. Introduction: Systemic Issues

The articles from Statistics New Zealand, NASS, RTI and Statistics Netherlands all focus on efficiency gains through the integration of production systems. In most cases, the re-engineering of the systems that is necessary for integration leads to quality improvements through greater consistency and timeliness of statistical outputs. The systems also aim to give greater control to survey managers and the flexibility they need to optimise their processes. As a result, overall quality is better managed, opening the way for Adaptive Total Design as described by Eltinge et al.

Integration requires an overall framework (sometimes referred to as an architecture) for organising production systems. For methodologists and survey managers, architecture is a new concept which is worth close examination. The question "What is in an architecture?" is not easy to answer. What kind of new methodological issues does this raise? How is the concept of quality related to the concept of architecture, beyond the traditional quality of a survey or of a system designed to process information? Moreover, as the Statistics Netherlands article asks, how do we get from the concept of total survey design to total network design? Does this question our traditional definition of quality, which so far has been very much linked to a single data collection rather than to a set of data collections?

We will reflect on those issues through an analysis of the work carried out in different fora and described in the articles produced by Statistics New Zealand, NASS, RTI and Statistics Netherlands. Our approach is compatible with, and complements, Eltinge et al.,

[1] Eurostat, Directorate for Corporate Statistical and IT Services, Bâtiment Jean Monnet, L-2920 Luxembourg, LUXEMBURG Emails: Daniel.Defays@ec.europa.eu and Jean-Marc.Museux@ec.europa.eu
Acknowledgment: Authors wish to thank Walter Radermacher, Chief Statistician of the European Statistical System, for support and inspiring comments.

who proposes a framework for analysing systemic issues. We lay greater emphasis on a limited set of topics: the international perspective, the constraints imposed by architectures and standards, the problem of design/specification raised by modular approaches to data processing (called the "frame problem" in the rest of this discussion), some new attributes of architectures, and issues of implementation and data access.

3. The International Context

The integration of production processes belongs to a new set of initiatives aimed at industrialising the production of official statistics. This is acknowledged in Eltinge et al. among others, with a reference to international frameworks such as GSBPM (Generic Statistical Business Process Model) and GSIM (Generic Statistical Information Model). The business case for this kind of endeavour is clear and has been described several times: at the level of UNECE (United Nations Economic Commission for Europe) through the vision and strategic papers of a high-level group on business architecture for statistics (now renamed the highlevel group for modernisation of statistical production and services) (Statistics Netherlands 2011); at the EU level through a joint strategy developed to improve the efficiency of production and quality of European statistics (Eurostat 2010); and at national level, as illustrated in the articles under discussion as well as in countries such as Australia (Studman 2010), Canada (Doherty 2011), or Sweden (Axelson et al. 2011). Among the arguments put forward in favour of an integrated approach, there are the constraints put on resources which require greater production efficiency. This is recognised in all the articles in this issue as one of the main drivers of change. There are also users' expectations regarding the rapid delivery of information, or new data needs. Statistics New Zealand, for instance, has developed an environment where there is room for experimentation in testing scenarios and rapidly designing new indicators. Then there is globalisation, which requires the merging of different kinds of data from different origins, by improving the interoperability of national systems for example. As regards the proliferation of data sources, in some countries surveys are now used to complement administrative sources. There are also quality improvements brought about by the use of paradata to handle nonresponses better (RTI International) or to facilitate the use of mixed-mode data collection (NASS). A more complete list of drivers can be found in Radermacher and Hahn (2011).

4. The Concept of Architecture

Architecture is an abstract concept used in different contexts. All articles refer to it as an appropriate framework to deal with system complexity. This section explores some of the dimensions of this concept building on the articles contributions. The ISO/IEC defines an architecture as "the organisation of a system embodied in its components, their relationship to each other and to the environment and the principles governing its design". Eltinge et al. propose to qualify architecture through "the performance characteristics of the components, the integrated system, and system outputs". An architecture is a model for the design of business processes and related supporting information systems providing building blocks, principles and constraints aiming at facilitating the maintenance and the evolution of complex and open systems. An architecture encapsulates, *ab initio*, some

knowledge and/or assumptions regarding the functioning and the organisation of the system. An architecture aims at striking the right balance between flexibility, rationalisation/efficiency, and quality. Information can be injected *a priori* in two different ways: either it is encoded in the models through parameters, input files, orchestration of building blocks or it is imbedded in the system itself. For instance, a building block can cover different GSBPM steps (say, A and B). In that case, the information that step A has to be followed by step B is encapsulated in the system/architecture and cannot be modified by the user. All architectures referred to in the articles build on the principle of modularity as the expression of the need for flexibility. The processing of information should be split into different steps corresponding to the different modules of the system. In the Statistics New Zealand article, statistical processing is carried out as an assembly of generic modules. NASS excludes multifunctional applications. Statistics Netherlands places modularity at the centre of its new system and RTI models a data collection as a combination of actions and steps, implementing those sequences in a standardised way to make them reusable. Eltinge et al. presents the increasingly used GSBPM framework as leading to highly modular systems. However, in doing so, the elementary blocks (representing operations, modules, GSBPM steps) out of which the processes are constituted represent the words of a language used to model statistical processing. Statistics Netherlands rightly questions the granularity of the modules, which, according to our analogy, means the vocabulary used to define processes. Eltinge et al. underline that in some cases standardisation may "prevent a given survey from making optimal use of survey-specific information" or "preclude some types of local optimisation of a methodological design". Regarding how the modules are defined, the articles do not go into sufficient detail, but it is clear that systems which lean on primitives such as exponential smoothing or linear interpolation or primitives such as historic imputation, seasonal adjustment (with a prescribed method encapsulated in the primitive), or micro aggregation (where the method for defining the micro aggregates from the original data is fixed) will have different "expressibilities". As defined for instance in Artificial Intelligence (see, for instance, Finlay and Dix 1996), an expressive system will be able to handle different types and levels of granularity of requirements. The more complex data processing it will enable to implement, the more expressive it will be. The range can span from a large set of statistical macros to a limited number of modules aligned to the GSBPM steps. A large number of modules gives expressibility and flexibility but makes the description of a process look like a computer programme. A system based on a limited set of modules will be more rigid and it will mean that a number of specific methods cannot be run. Nevertheless, it will probably lead to more standardised processes, which are easier to manage. There is no doubt that this is a key issue which should be carefully addressed with the right skills. In this context, it is important to draw on the competence of methodologists.

Similar issues arise when standards to represent the data and the metadata (including paradata) are embedded in the architecture. Here, too, the structure will restrict what can be represented. In SDMX (statistical data and metadata exchange), for instance, the notion of variable or question does not exist, contrary to what is offered by the DDI (Data Documentation Initiative). The articles under discussion do not go into much detail on how to represent key information. At European level, technical standards are currently

being developed to structure, reference and process metadata (ESMS – Euro-SDMX Metadata Structure) and quality reports (ESQRS – ESS Standard for Quality Reports structure). Although like the modules they are extremely useful, they restrict what can be expressed. It would be interesting to explore if similar limitations exist in the way paradata are represented in Nirvana – the integrated data collection system designed by RTI.

5. The Frame Problem

Another difficult issue linked to the concept of processing as an assembly of modules is the implicit assumption that operations and data/metadata are independent of each other. More explicitly, a module is supposed to accept as input a set of information related to the data, metadata and parameters relevant for the processing to process the data and deliver an output which can be transformed into data, new metadata, or, for instance, a report on the processing. But as the processing is modular, in most cases the output in itself has no other *raison d'être* than to feed the next step of the process. But which step? If the modules are designed independently, and this is necessary in a modular approach, any kind of step which makes sense from a statistical or logical viewpoint is possible (imputation after validation, secondary cell suppression after the masking of primary confidential cells, etc.). However, how can we be certain that the output will be self-contained and ensure that all the relevant information for any kind of subsequent processing is available?

The traditional system from which GSBPM is derived provides an overview of the whole chain and the information needed at the different steps of the processing. In a modular and network approach, this is no longer the case. To take a simple example: when a value is imputed in an edit module, how should attention be drawn to it? By flagging the value as estimated? It may be that later on it will be important to know that this was estimated using a linear interpolation technique in order not to conclude to a linear trend which will be detected by an analysis module. You may well need to know even more about the method which has been used. Everything could be relevant *a priori*.

However, keeping the metadata/paradata of all information that might become relevant at some stage of a potential treatment is just not an option. This issue has similar features to the well-known "frame" problem in Artificial Intelligence. Here, the frame problem is the challenge of representing the effects of actions without having to represent explicitly a large number of intuitively obvious noneffects (see for instance Callan 2003).

6. The Quality of an Architecture

Methodologists are used to design surveys to optimise their overall quality under constraints (see for instance the interesting framework proposed by Eltinge et al.). This includes reducing the effects of nonresponse, frame coverage error, measurement and data processing errors while at the same time maximising timeliness and relevance. Often, other quality factors such as comparability and consistency are also taken into account (see for instance the Eurostat definition of quality). Strictly speaking, the last components are not characteristic of a survey, as they refer to the relationship which the statistics may have with statistics issued from other sources. They are, in fact, the property of a system, the kind of entity which becomes highly relevant when surveys are integrated. The shift in focus from a single survey to a set of integrated surveys affects the way quality has to be

defined. In this new context, comprehensiveness (see the Dutch experience), redundancy, and local versus global optimisation become relevant issues. When the architecture is at stake, quality issues go beyond the quality of the system of integrated data collections it supports, as recognised in Eltinge et al.. Capacity is also relevant: to host new surveys (extensibility and adaptability); to promote reuse of components and interoperability (efficiency); to use standards; to be consistent with frameworks (such as accounting frameworks); to exploit the same set of data for multiple purposes; and to rationalise data processing in a sustainable way. This is an area which is only just starting to be explored by statisticians.

7. Process Standardisation Issues

In the model designed by Statistics New Zealand to meet the need to create or improve statistical output in a flexible way and make it a baseline activity for statistical subject-matter specialists, the common methods used to process the data (editing, outlier detection, imputation, adjustment of weights, estimation of model parameters, etc.) are combined into an end-to-end production process known as a configuration. In the Common Reference Environment (ESSnet CORE 2012), a proof-of-concept project carried out at European level to test the feasibility of modelling and orchestrating a statistical process as an assembly of services, work flow is also a key concept. When processes have to be integrated or reused, they need formal representations that can be processed by machines or by statisticians. Of course the notion of process is not new for official statistics and there are numerous methods for designing, managing and documenting them (see for instance the EPMS standard currently being drawn up at European level; Eurostat 2013). Nevertheless, the representation of processes as objects which drive processes as used in CORE and in Statistics New Zealand deserves to be highlighted. In the Dutch experience, the concept is slightly different, resting on the notion of a set of stable datasets (called steady states) and a production chain for the exchange of steady states. This goes beyond the notion of production lines and again only makes sense as part of a holistic approach.

More systemic approaches in statistical surveys and data collections will lead to new kinds of information concepts and objects that still have to be precisely defined and supported in production systems. These aspects were missing from GSBPM, but they have been partially integrated into the recently released Generic Statistical Information Model.

As stressed in Radermacher and Hahn (2011), "looking at individual production processes and grouping them into larger entities is not enough. All processes will need to be tackled in a holistic approach with a view to enabling their integration and modularity within the whole statistical authority, across statistical authorities and with international organisations. Thus, there is a need to enlarge the scope of standardisation with meta-processes – going beyond the GSBPM – that allow for a bird's eye view on production". Figure 1 illustrates the scope of a statistical "factory":

In this broader context, an architecture can be extended to cover nonproduction processes such as common professional standards for statistics as developed by the European Statistical System project of an European master programme of official statistics, standardised press releases and other press activities, and so on.

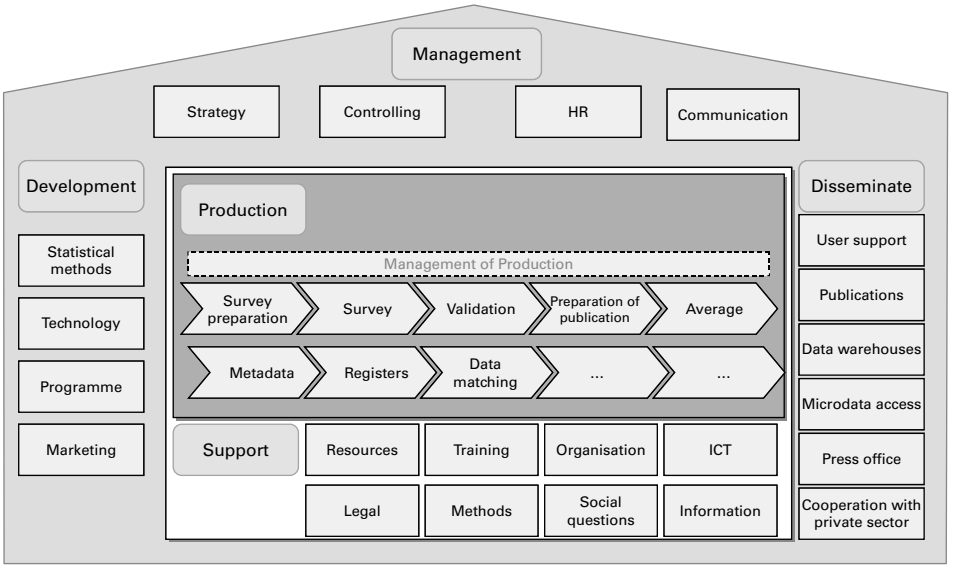

Fig. 1. The statistical factory (Radermacher and Hahn 2011)

Finally, the standardisation of processes/metaprocesses cannot be disentangled from a "product"- component, which is the second strand developed in Radermacher and Hahn (2011). As long as we define our products in a tailor-made way, fit for (and only for) the purpose of specific users, the industrialisation-potential of processes will be extremely limited. The new architectures for Official Statistics should include the building blocks for "multipurpose-products", which are ideally "*prêt-à-porter*", meaning of high quality, usable for many purposes, and provided in the form of a coherent set of mesodata organised through accounting frameworks.

8. The Implementation Issue

Although the industrialisation of statistical processes and the rationalisation of information systems have been discussed for several years, success stories and implementation reports are still rare. On the basis of the four articles, good practices can be derived for the re-engineering production systems.

All the initiatives target a paradigm shift in the production environment. This requires high-level skills, dedicated governance mechanisms, an adequate mobilisation of resources and deep involvement on the part of the organisation business process owners.

All transformation initiatives include well-defined governance that breaks down hierarchical barriers. Architecture boards are frequently required for translating strategic goals into models, operationalising the goals and balancing local and corporate interests.

The ongoing experiences described in the articles show that while the final strategic goal is very important, it is nonetheless essential for the programme to be structured as a sequence of transitions towards intermediate states, where benefit can be demonstrated and reorientation is possible should there be changes in the environment. Success stories

are frequently built on past failures which have been analysed and served to build capabilities.

Resource allocation must be agile as it can be necessary to reallocate resources (be it financial or human) rapidly at any stage of the implementation. This is difficult to integrate in a standard cost-benefit analysis. A recommended way (Statistics New Zealand) is to proceed step by step (phasing), measuring the benefit of specific pilots, either further optimising the future state or continuing the planned implementation.

The articles highlight the need for the strong involvement of users. Agile development involving users, developers, and designers is also proposed as a means for the gradual creation of competence.

The skills for designing a new system are not always available in house. In the event of outsourcing, shadowed capabilities (RTI International) have been found to be necessary in order to build in-house competence to ensure the sustainability of the system.

Eurostat has recently designed a programme for developing a set of common infrastructures to support the implementation of the ESS vision for the production of EU statistics (Eurostat 2009). The main technical infrastructure components include: 1) a secured network supporting the flow of information among ESS partners; 2) a platform for sharing common services; 3) common data repositories; and 4) a common validation architecture. The approach combines the business realisation of key flagship initiatives which will demonstrate and pilot key features of the ESS Vision (exchange of information and interoperability of statistical processes) while gradually building up the infrastructure components of the future system. The approach involves a paradigm shift from a stovepipe, bottom-up and project-based approach to a programme-based approach closely tying together business outcomes and crosscutting infrastructure developments.

The programme infrastructure being set up builds on the above principles adapted to the international and decentralised context. For more information on current developments, see Eurostat (2012).

9. The Data Access Issue

In a systemic context, as acknowledged in the business case for official statistics in a global information market (Radermacher and Hahn 2011), access to detailed and confidential information by partners or external users has to be integrated into the architecture design. Official statistics have traditionally maintained a privileged relationship with information providers. Indeed, the protection of individual information collected for statistical purposes is usually seen as the foundation for the quality of the primary information collected. This has led to a well-developed security and statistical disclosure limitation architecture components. However, it is also building barriers against the wider integration of statistical systems and the necessary specialisation of statistical work. The move towards reusing existing information (currently coming mainly from administrative sources, as shown by the Statistics New Zealand article) is creating a slight shift in the focus of statistical confidentiality, to the extent that there will be less direct interaction between statisticians and data providers, although the principle of respect for confidentiality will have to remain enshrined in the future architecture. Given the place of data analytics in the new architecture, innovative solutions must be found to ensure the

appropriate trade-off between flexible access to detailed data and statistical disclosure limitation. The Westat/NCHS article gives a detailed account of the design of an online analytic system targeting external users, which is required in order to cope with statistical confidentiality. In an integrated system such as the European statistical system, new architecture should push the provision for security towards the system's border, implementing the Schengen model for the movement of individuals within the EU (see for instance Bujnowska and Museux (2011) on the implementation of the Schengen model for research into microdata access). Under the Schengen approach, all the data collected under European legislation should be considered as the common good of the European Statistical System and, consequently, each national authority should be authorised to access and grant access to all European data. This could only be possible if a common architecture defines rights and duties, protocols and basic principles (most likely to be enshrined in the appropriate legislation).

In the medium to long term, the public's experience of statistical confidentiality is expected to change, given the ever-increasing amount of individual information deliberately or unconsciously made available on the internet through social networks or mobile devices. Statistical confidentiality will no doubt have to be redefined to take a changing reality into account. A new paradigm will almost certainly have to be found.

Efficiency gains, quality improvements, and higher reactivity push statisticians to conceive their systems in a more holistic way. This requires new concepts; this raises new issues. The shift from systems to architecture, from local optimisation to global optimisation, from the production chain to the factory is on its way. The current empirical approach needs to be embedded in a more theoretical framework. This is a new challenge for statisticians.

10. References

Axelson, M., Engdahl, J., Fossan, Y., Holm, E., Jansson, I., Lorenc, B., and Lundell, L.G. (2011). Enterprise Architecture Work at Statistics Sweden. Paper presented at the 2011 International Methodology Symposium, Statistics Canada.

Bujnowska, A. and Museux, J.-M. (2011). The Future of Access to EU Confidential Data for Scientific Purpose, 2011 UNECE-ESTAT Work Session on Statistical Data Confidentiality.

Callan, R. (2003). Artificial Intelligence. Basingstoke: Palgrave Macmillan.

Doherty, K. (2011). Update on How Business Architecture Renewal is Changing IT at Statistics Canada. Working paper 3, MSIS 2011, United Nations Economic Commission for Europe.

ESSnet CORE (2012). Common Reference Environment, EU Commission funded project. Available at http://www.cros-portal.eu/content/core-0 (accessed February 14, 2013).

Eurostat (2009). Communication 404/2009 of the Commission and the European Parliament on the "Production of EU statistics: a vision for the next decade". Available at http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2009:0404:FIN:EN:PDF.

Eurostat (2010). Communication from the Commission on the European Parliament and the Council on the Production Method for EU Statistics — Joint Strategy Paper, 5th Meeting of European Statistical System Committee, document ESSC 2010/05/6/EN.

Eurostat (2012). The ESS VIP Programme, 15th Meeting of European Statistical System Committee, document ESSC 2012/15/6/EN.

Eurostat (2013). Better Documenting Statistical Business Processes: the Euro Process Metadata Structure, forthcoming paper METIS UNECE Work Session, May 2013.

Finlay, J. and Dix, A. (1996). An Introduction to Artificial Intelligence. London: UC Press.

Radermacher, W. and Hahn, M. (2011). The Business Case for 21st Century Official Statistics. Unpublished paper, available at http://www.cros-portal.eu/content/business-case-21st-century-official-statistics (accessed February 19, 2013).

Statistics Netherlands (2011). Strategic Vision of the High-Level Group for Strategic Developments in Business Architecture in Statistics. Paper presented at the 59th plenary session of the conference of European statisticians, UNECE.

Studman, B. (2010). A Collaborative Development Approach to Agile Statistical Processing Architecture — Australian Bureau of Statistics (ABS) Experience and Aspirations. Working paper 3, MSIS 2010, UNECE.