Building a common production environment – the Swedish experience

Johan Erikson, Martin Odencrants[1]

Abstract

Statistics Sweden is on an ongoing journey to use more standardized and process oriented IT systems and methods in the statistical production. We call this building a common production environment, where our IT tools are combined to create an integrated support for the production from data collection through process and analysis to dissemination and communication. The system is built to be flexible based on the design choices made by each survey, while keeping a standardization perspective in mind. Metadata will be used to steer the production process, using GSIM as a corner stone for this information. The system will have three main parts: Data collection which is already much in place through the Triton system, Process and analysis using a new system called KLON and Dissemination and communication where presentation in our statistical databases on our website will be the first component. A sharepoint dashboard within our process support system will be used to set up and describe surveys, their design and their production flow, to set up the necessary metadata and to run and monitor the production process. While the creation of this system will be ongoing for several more years, in 2016 we will have surveys using all three parts. Triton is already used by more than 30 surveys with 15 more coming in before april 2016, KLON is in production and will have around 10 surveys in 2016, while the new dissemination database and the connection to the statistical databases will be implemented in 2016 as well. The presentation will give an overview of the development so far, the contents at the time of presentation and future plans, as well as discuss some lessons learned so far.

Key Words: standardization, new technologies

1 Introduction

Statistics Sweden (SCB) has for, for several years, been working towards a more standardized and process oriented approach to statistics production. An important element in this ambition is to build and establish a common production environment. The foundation for this environment is an SCB adapted version of the GSBPM model where metadata, using GSIM as a basis, is used to configure and steer the production process for each survey in the system. The environment will contain IT-tools and information on routines and best practices to use for process steps ranging from data collection to dissemination of data. IT-tools for performing activities related to each step, e.g. estimation, are combined in a flexible way in order for the system to handle the shifting needs that surveys at SCB have. By consolidating tools, routines and best practices in the common production environment the need for unmotivated process variation decreases as well as the need for product specific IT-tools. As a result the common production environment is also expected to simplify maintenance of tools and activities related to the production process at SCB.

[1] Johan Erikson, Statistics Sweden, SE-701 89 Örebro, Sweden, email:. Martin Odencrants, email: martin.odencrants@scb.se

Supporting a process oriented approach to producing statistics is one purpose of the common production environment project. Other aspects that play an important role in the development are improved traceability and reproducibility as well as increased control over both micro- and macro data.

The project to build and establish a common production environment will continue for several years. At this point the system consists of three modules connected to a sharepoint dashboard within our process support system. Activities connected to data collection are contained within the Triton system, KLON combines tools used during the data processing and analysis phase and tools used for dissemination of data are contained within a third module. Eventually, the common production environment will span the full width of the production process in a seamless way.

This paper will describe the corner stones of the common production environment project and aspects that give restrictions that development need to take into consideration. We will discuss the process oriented approach to statistics production from data collection to dissemination and the lessons learned so far in SCB:s ambition to standardize the production process and improve traceability, reproducibility and content coordination.

2 The road to a common production environment

Like many statistical offices around the world, Statistics Sweden have a number of common tools for different parts of the production process, some of them are used by many surveys and others are more specialised for parts of the production process that are not necessary for all surveys. However, these tools are not automatically connected or integrated. Many surveys have survey specific systems and also specific programs or scripts written in SAS or SQL that are used in combination with the common tools. Each survey has also stored its own data using very different structures of this storage, meaning that combining data from several surveys needs a lot of manual treatment, matching and restructuring of data. Over the last ten years many survey specific systems have become old and in need of replacement while the maintenance of systems and integrations has become complicated and expensive. With a number of new tools being developed or planned around ten years ago, like SELEKT for selective editing and Prisma for manual coding, the need of a better IT environment was apparent. At the same time, Statistics Sweden decided to move to a process oriented view on statistical production, with more standardised ways of running the production across surveys. These were the main reasons to start the development of a Common production environment that integrates all the common tools with the goal of a seamless system of IT components that together can be used to run the statistical production.

The development of the common production environment is taking a stepwise approach. It was decided to start with data collection for business surveys since this process had the most prominent needs. This development started in 2009. From 2012, a development of a component for processing and analysis was started with the specific first goal of supporting a few specific business surveys in economic statistics, while at the same time work started to replace old systems for dissemination. This means that the common production has three different main parts, while the fourth but not least of the parts is common metadata and structures that are overarching the other three parts, more emphasis on this has been given from 2015. Implementation in actual surveys is on-going at the same time as further development, meaning that more complex surveys are not implemented first.

The goals that Statistics Sweden hope to achieve with a common production environment are:

- Standardisation of how the statistical production process is run. Building a new environment will mean that we can implement more standardised routines across surveys.
- More emphasis on design, design choices steer the production. We aim for a design-based metadata driven system where the design choices will give you access to relevant tools, templates and support, and data flows between services or IT tools are also decided by design choices.
- More efficient production – when surveys work in the new systems, it should be possible to put more emphasis on working with design and content rather than IT systems. A more efficient production should free time for other tasks such as developing new products and catering more to user needs.
- Quality assurance of the statistical production will be built in, so that standard approaches and demands are pushed out to the surveys, access to our Process support system will be better in the new environment and surveys will get relevant information automatically based on their design choices.
- Traceability (in data and programs) and possibility to reproduce statistics will be built-in features of the environment. These features guarantees that Statistics Sweden satisfies outside demands for this.
- Less manual treatment of data. With automatic flows between services and tools manual work on preparing and matching data will be minimised, this will also free time for other tasks such as design work.
- More efficient maintenance of IT systems. Even if the new systems are bigger and have many users, the maintenance can much easier be controlled centrally, meaning less risk of systems getting technically old or outdated. Less dependency on single persons will also be possible with maintenance teams based on the new systems.
- Prepare for coordination of surveys. Coherent statistics is one of the most important goals of Statistics Sweden´s Strategy 2020. Moving to coordinated surveys will not in itself come out of using a common production environment, but the new environment will make such work much easier with data organised in a much more structured and standardised way than today. This will also enable easier re-use of data.
- Easier to build new common tools when data structures and metadata are standardised and coordinated.
- Easier to set up new surveys when they can be configured in an existing system rather than having to build a new one.
- Possibility to implement new and improved tools in one place. New features and tools can be "rolled out" within the common production environment and be accessible to all surveys immediately. Today each development has to be implemented in a number of different systems.

3 The common production environment and its parts

As we see it now, the common production environment consists of four parts; Data collection, Processing and analysis, Dissemination and Common metadata tools. The first three are areas that are distinguishable within the statistical production process and that can operate independently, yet benefit from being connected. The connections are generated by having common metadata tools and automated data flows.

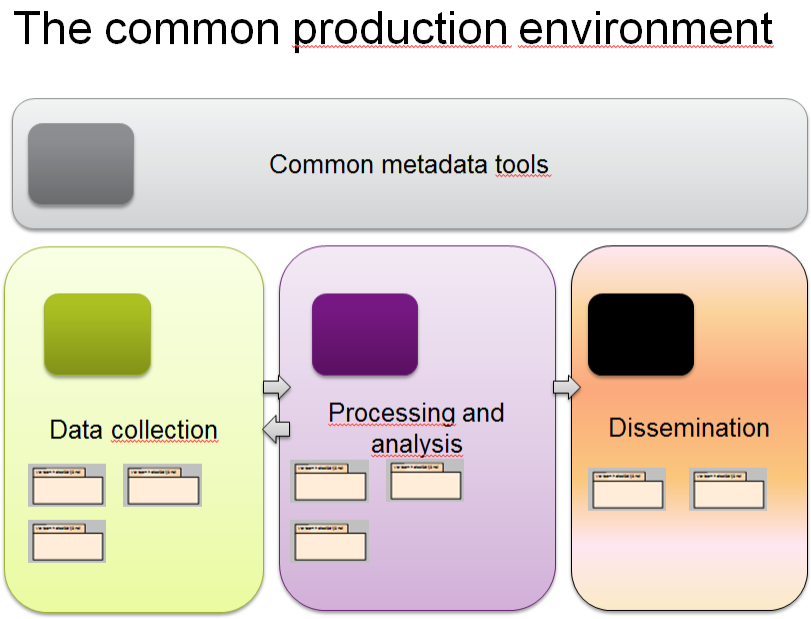

Figure 1: The common production environment

The different parts in turn consist of a number of logical services. For data collection these go under the common name of Triton but can in fact be distinguished as services for web data collection, paper form data collection (optical scanning), distribution, treatment of objects and micro data, error signalling et cetera. For Processing and analysis there are two main services; calculations (KLON) and macro editing (Veritas). For dissemination there are services for planned disseminations (REDA), a service for publishing of tables on the website and other services envisioned are delivery to customers and deliveries to international organisations. Over time, our different common IT tools are integrated into the common production environment and developed into services. Planned integrations apart from those mentioned for dissemination are sampling, interview data collection et cetera.

The common metadata tools are also logical services that serve and steer the other parts of the production environment. Services that we have or envision here arethe already existing Process areas (dashboard) and Statistical content (MetaPlus), a database on surveys, a service on contracts for data deliveries (from other authorities as well as other surveys and also respondents that send data files) and a common login service for both respondents and customers.

3.1 Data collection - Triton

Triton is the name for a number of services and tools covering Data collection and Micro editing. The first version was released in 2011 (Erikson 2011), and since then continuous further development has taken place. The development of Triton also lay the foundation stone for some common functionality that will expand to the other parts of the common production environment as well, such as the “dashboard” and the activity list quality assurance system. Today the Triton “system” has the following parts:

- "Dashboard" with access to all tools and services. A survey specific and also collection round specific dashboard is created based on the design choices that the survey makes. From the dashboard, you can access all the other tools and services.
- Paradata reports on data collection progress. These are presented in the dashboard.
- Web data collection and paper form data collection. Since business surveys and public sector surveys have been the priority this far, interview data collection is not yet included but planned for 2017 or 2018.
- System for defining and setting up collection variables and value sets.
- System for sending out letters and e-mails to respondents. This will be expanded with penalty letters and so on in 2017 or 2018.
- Functionality to load a drawn sample. A new sample selection service has been created in 2015, and this will be integrated into the Triton part of the common production environment in 2016.
- Data storage with traceability of all changes made to data.
- Signal, which is a service for error localisation. Signal can use both traditional editing and selective editing. If selective editing is used, our SELEKT tool is used and integrated. Edit rules can either be created in our web collection tool and exported to be run in Signal as well (the first choice) or, if that is not possible, by writing specific edit rules scripts that can be run by Signal.
- EDIT which is the main tool for treating reporting units. The EDIT tool is the one that is expanded most over time, since implementing new surveys in the system pose additional demands on the needs to look at data and units in different ways.
- Automated communication and data transport services between all integrated tools, so that no manual transport or treatment is necessary.
- Treatment of duplicates if respondents send data more than once.

3.2 Data processing and analysis - KLON

KLON is designed to handle calculations related to the process steps contained within Data process and Analyse in GSBPM. The environment consists of three interacting parts. A user interface, data storage and SAS code to process data. Traceability, reproducibility and data coordination were fundamental requirements when the architecture was designed. Another key feature of KLON is the use of metadata to configure and steer data processing. With KLON, more emphasis is put on survey design and choices made in the design stage are used to steer the production with help of metadata.

Subject matter specialists and methodologists are the two main user groups of the KLON system. The demands from each group differ and there is support for each competence to work in the system. Subject matter specialists execute SAS code in KLON by defining which time period and data processing step the code relates to.

SAS has been the main tool used for data processing at SCB for many years and is also the tool used to process data in KLON. Survey methodologists mainly do their work in SAS where they write SAS code for data processing, update the SAS code when needed, develop new code to produce new statistics etc. A current example of the latter is the development of production value indices. All services in KLON are coded in SAS. Calculations in KLON are, as far as possible, made with common SAS tools, e.g. ETOS for estimation and BANFF for imputation. A central question related to services in

KLON is how generic the services can be and how survey specific they need to be. At this point there are examples of services that are generic, e.g. seasonal adjustment. Estimation is an example of the opposite, estimation at this point is survey specific.

Traditionally, at SCB, tasks related to the production of statistics and investigative work has been made in the same environment. In KLON there are separate environments for production, test and development. Tasks related to survey development cannot be performed in the production environment, these tasks needs to be performed in the development environment, since access to production code and data are restricted. At its core, performing different tasks in different environments is a good thing even if it also increases costs for maintenance and increases resources needed for development.

At this point, ten products use KLON in production and a few more surveys are planned to enter the environment in the near future. The system is still very much under development and there are several challenges that need to be addressed before the system is ready for a larger number of surveys. These challenges are related to data storage, the different environments, work procedures etc. Still, SCB has a plan for how the environment should be developed in order to handle the challenges that have been identified.

A current issue is incorporating a new tool for macro editing, Veritas. Veritas is built with SAS visual analytics, importing data from KLON. If possible errors in micra data are found in Veritas, possible changes will be made in EDIT so that we keep full coherence between data in all parts of the production environment.

3.3 Dissemination

The development of dissemination tools started as a continuation of a plan to update the old tools since they needed to be replaced with modern technique. This was around 2013 when the development of the common production environment was already on its way. It was decided to start merging these initiatives so that new tools would be developed but offered as services within the common production environment. A first part was to replace the application for describing tables on the Statistics Sweden website. This was done in 2014 outside the common production environment. In order to integrate that application and others in the dissemination area the central tool for planning and describing disseminations needed replacing first. That work started late 2014 and has just been completed, the new database REDA was released on April 29 2016. It will be used by all surveys for dissemination in 2017 onwards. With this database in place, it is possible to replace the remaining tools for publication on the website. That work will start in the autumn of 2016. After that, work will be expanded to other dissemination forms like deliveries to customers and international organisations. A central part in the further development will also be to integrate the data flow from finished calculations (statistics ready to be published) and actual dissemination. This flow still involves several manual parts.

3.4 Metadata tools

In order to have a seamless integration between the three production parts (data collection, processing and analysis and dissemination) the common production environment needs several common metadata services to connect the different parts. Statistics Sweden has dedided to build these metadata tools to align with the Generic Statistical Information Model, GSIM (UNECE 2013a). The GSIM model has been developed by the UNECE to complement the Generic Statistical Business Process Model,

GSBPM (UNECE 2013b). In short, the GSPBM is a way to describe the process steps in the statistical production process, and GSIM describes the information that is needed to run the production process.

GSIM can be broken down into four different parts:

- Concepts, which can be said to cover what has been described as metadata for a long time, including definitions of statistical concepts, variables, value sets, classifications, object types and populations.
- Structures, which describes the actual data layers, how data is stored in different tables, data sets et cetera.
- Business, which covers metadata on how surveys are structured, which process steps, methods and services they use and how the production flow is designed.
- Exchange, which covers what information is gathered/collected from whom and what information is delivered to whom.

Together, these four parts cover all the necessary metadata to run the statistical production process, and also allows this metadata to be used in an active way actually steering the production. To make the Swedish common production environment metadata driven, we envision to have common metadata tools for all these parts. These are not all in place yet, so the three existing production parts have some temporary specific solutions that will have to be adjusted when the common tools are in place. It should also be made clear that metadata exist on three different levels of detail in the production environment:

- Common metadata that is used by several parts of the production system or that is used to steer the production flow. This metadata is held centrally. This can be for example the surveys we have, how they are organized and their design choices, the contents and metadata describing deliveries between surveys.
- Metadata that is used in only one of the three production parts of the system, but is used by several services in that area. This can be for example data structures within the production flow. These are stored centrally within each of the three parts.
- Metadata that is unique for a specific service (e.g. survey specific settings within a tool) are only stored within that tool. This might be for example settings on how to present data to staff for editing and so on.

The system that we already have in place in place is Metaplus, which covers the concepts part of GSIM. Since Metaplus has been in place for a number of years, it is not fully aligned with forming an integral part of a full metadata system in line with GSIM, so an updated version of Metaplus is in the planning stage.

The metadata in Metaplus needs to be complemented by other common metadata in alignment with the GSIM model. Work has started on two additional metadata tools. The first covers the business part of GSIM and contains is a database over surveys and their production rounds, process steps, status codes and other process steering metadata. The other is metadata on data deliveries; between surveys and from outside of Statistics Sweden to surveys or registers. Development of both of these started in 2016 with a first version being released by the end of the year. Development will continue into 2017.

4 Current and future plans

As has been shown above, the status of the different parts of the common production environment varies. Triton had its first version in 2011, KLON in 2015 and REDA will be released in 2016. Other parts are still under development and some are on the map but not yet started. Therefore, it is envisioned that the development will continue until 2020 at least before we have a stable version of all parts of the environment. For data collection we plan to expand to surveys of individuals and households as well, and that means integrating systems we have and systems that might change in the next few years. Our systems for web data collections and telephone interviews need to be replaces, and we're looking into Blaise as a collection instrument. Expansion of file transfer and machine to machine-solutions for data provision are also envisioned. For calculations and output editing lots still has to be done, especially to get output editing to be an integral part of production. For dissemination other means of dissemination than through the statistical databases will have to be integrated. And the development of common metadata solutions to hold the different parts together needs to continue. The dashboard functionality in Triton is also planned to expand to all parts of the production process.

5 Lessons learned

The development of the common production environment has been on-going for a number of years now, and there are a number of lessons that we have learned so far:

Continuous implementation means both pros and cons. The pros are that you can start with simpler features and expand to more complex versions over time. But on the other hand the period when you have two maintain two production environments becomes longer. Sometimes you cannot phase out old systems because the new systems are not complete.

We have done this in steps, building in different parts first and combining them later. This probably means that the development of each part is easier at the beginning since you work with only one area at a time, but on the other hand fitting them together at a later stage means new problems since they were designed in different ways. We still think the stepwise approach has worked rather well.

There is a need for refactoring over time, especially if you work with agile development. Solutions will be built in stages, where version 1 is not always the final or best one. This is something that needs to be understood by stakeholders, and resources need to be available for rebuilding already existing stuff as well as adding new functionality.

Plans and priorities need to be transparent to everybody, otherwise disappointment will follow. There is a need to balance demands from current users and new users, and everybody needs to understand that resources and time are limited and not everybody can get everything they want at the same time.

The process of change is much more than building and implementing new systems. New ways of working are even harder to implement. Therefore, a strong and continuous support from management is necessary, both from top management and managers whose surveys are implementing the new tools and ways of working. Surveys at SCB have a history of producing statistics within their own custom made system. Moving towards a common environment will decrease the number custom built systems. For a specific

survey this means that when entering the common environment they will have to adapt their way of working. The overall purpose of the common environment needs to be clear when a survey are to be implemented in the new environment. The purpose of the new environment is not to incorporate all features that where available in the old system. Trying to clone old production systems in order for new surveys in the system to be happy will make the new environment difficult to maintain and it also takes a lot of resources to adapt the environment before a new survey can enter and start using the system.

Our experience is that there is a risk of developing stove-pipes between systems/processes instead of surveys. Therefore it is necessary to dedicate resources to the holistic approach, the overarching architecture and the glue between processes and systems.

Building a common production environment for the statistical production process as a whole is an ambitious project. The size and complexity of the task makes it impossible to foresee all demands the system has to live up to. System support has to be designed to meet the needs from different competences, the varying conditions different surveys face as well as taking overall SCB demands into account. SCB has chosen to use an agile approach when developing the common production environment.

This means that the needs that different competences have will be met over time, the system as a whole is developed in an incremental an iterative manner. New system features will be added to the system continuously.

It is important to cater for the needs of all the main user groups. In KLON, the majority of the resources for development have been targeted towards the needs of subject matter specialists. As a result; features that methodologists need in order to work efficiently still need to be built. The restrictions that the system puts on how you can work in the system as a methodologist has raised questions regarding the KLON system from a methodologist perspective. In order to communicate how development is prioritized and the strategy of development; there needs to be clear channel where all interested parties can gather. And there also has to be a way for all competences to communicate their needs to the development team and feel that they are being listened to.

With different parts developed by different teams, there is a strong need for communication between teams. There are many interactions between systems. The dependencies between e.g. KLON and Triton is one example. Imputed values are delivered from KLON back to Triton and micro data are imported from Triton to KLON for further processing. Changes in the Triton architecture will in some cases affect features in KLON and vice versa. When changes are made in one system they need to be communicated to the team responsible for the other system.

Guaranteeing a stable system with high performance is more important than implementing new surveys. Pushing new surveys into the new environment before some basic features are in place will give the new environment a bad reputation that will be hard to repair.

There are many challenges involved in developing the new environment, large and small, related both to technical aspects and the way to work in the new system. Looking back choosing to develop the system incrementally and iterative has been a good decision. The

complex nature of the project makes it hard to foresee all needs a common production environment has to meet.

6 Conclusion

This paper has described the current state of SCB:s common production environment from data collection to dissemination and lessons learned. The project is still on-going and will continue until 2020 at least. When ready, SCB:s goal is a seamless production environment where all competences working in the system can find the tools they need in order to produce statistics of high quality. Several lessons has been learned along the way, two of the more important ones has been that guaranteeing a stable system with high performance is more important than increased implementation and that development should be made incrementally in cooperation with users of the system. At this point several key features of the common production environment (Triton, KLON and REDA)are in place, but much work still remains before SCB:s goal can be reached.

References

Erikson, Johan. Triton – a general tool for data collection and micro editing. Proceedings of Statistics Canada Symposium 2011.

UNECE. Generic Statistical Information Model (GSIM): Communication Paper for a General Statistical Audience version 1.1, December 2013
Available from:
http://www1.unece.org/stat/platform/display/metis/Generic+Statistical+Information+Model

UNECE. Generic Statistical Business Process Model GSBPM version 5.0, December 2013
Available from: http://www1.unece.org/stat/platform/display/GSBPM/GSBPM+v5.0