YTY - integrated production system for business statistics in Statistics Finland

Author: Antti Santaharju[1]

Abstract

In 2009 to 2014 Statistics Finland carried out a project, which planned and implemented an integrated production system (YTY) for the Business Register and 13 business statistics. The project was started to decrease the overlapping work in different statistical production systems. The same information was collected and processed in different processes, which led to decreasing coherence between business statistics. The maintenance of several different information systems consumed much resources. The YTY system was introduced in November 2013.

The YTY system includes solutions for the data collection phase. The reception of administrative data and direct data collections are carried out with standardised processes and tools. All input data are stored into the raw data warehouse. Data from different sources are integrated in the production database, where all the observations are edited only once for all the YTY statistics. The edited data are transferred to the data warehouse for finalised micro data, where information is re-used many times for the aggregates of several different statistics.

The YTY production system is managed by the process management system, which documents the planned and realised updates of source data and required edits. All automatic processing is performed by this process management system. Source data and their variables are described in the metadata.

The introduction of the YTY system led to the reorganisation of work processes and procedures between business statistics. Two production cycles for the reference years 2013 and 2014 have been performed with the YTY system and the benefits and experiences start to be realised. This paper presents these experiences and discusses how the introduction of the system has affected the statistical production and the resources used in it.

Key Words: Data warehouse, integrated production system, coherence

1. Introduction

In 2009, Statistics Finland launched a project with the objective to renew the production system of the Business Register. Simultaneously, the objective was that the Business Register would become the sampling frame of all business statistics. A lot of additional development pressure was also directed at the Business Register and business statistics, like the introduction of the enterprise unit that was looming in the near future and improving the coherence between business statistics. In this operating environment, Statistics Finland decided to expand the planned development objectives of the Business Register also to other central business statistics. Statistics Finland launched a project with

[1] Antti Santaharju, Statistics Finland, Mailing Address: FI-00022 Statistics Finland, email: antti.santaharju@stat.fi.

the aim to build an integrated production system for the Business Register and central business statistics.

The main objectives Statistics Finland wished to achieve with the integrated production system were:

- Coherence between the concepts of the Business Register and business statistics
- Common database and common variables for the Business Register and business statistics
- Converting data from old systems to the common database
- Harmonised production processes and common tools in statistics production
- Increasing efficiency in the time used in statistics production and to maintain the statistical system
- A common enterprise data warehouse, where business data can be selected on unit and aggregate level by users.

Statistics Finland launched the project to develop an information system in accordance with the above-described objectives. This work was carried out in the *Business Register into the core of data on enterprises* project that started in January 2009 and was completed in December 2014. The result of the project was the integrated production system for the Business Register and central business statistics called YTY. Saarikivi has described the background, organisation and implementation of the project in question in a paper published in connection with the ISI2013 conference[2].

2. Presentation of the YTY system

Statistics Finland's integrated production system for the Business Register and 13 business statistics (YTY) was taken into use in November 2013. At that time, the first production cycle to produce business data for the reference year 2013 started. The data of the Business Register and the 13 business statistics listed below were produced and released for the first time through the YTY system by the end of 2014 in accordance with the schedule presented below (excluding the Finnish affiliates abroad statistics, whose data pertaining to the reference year 2013 were published in April 2015). The data of the Business Register are produced in the YTY system with the schedule t+12. The business statistics produced through the system and their release lags are:

- Structural business and financial statement statistics (preliminary t+9, final t+12)
- Regional statistics on entrepreneurial activity (preliminary t+9, final t+12)
- Industrial output (t+11)
- Enterprise openings and closures (Q1:t-5, Q2:t-2, Q3:t+1, Q4:t+4,)
- Index of turnover in industry (t+75 days)
- Index of turnover of construction (t+75 days)
- Turnover of trade (preliminary t+22 and t+45 days, final t+75 days)
- Turnover of service industries (t+75 days)
- Wage and salary indices (t+45 days)
- International trade in services (preliminary t+5, final t+12)
- Foreign affiliates in Finland (FATS) (t+12)
- Finnish affiliates abroad (FATS) (t+16)
- Business Services Statistics (t+10)

[2] Saarikivi 2013: http://2013.isiproceedings.org/Files/IPS023-P3-S.pdf

Data editing for the statistics listed above takes place in the YTY production database, which is the work area of the persons participating in data processing. All extractions related to the dissemination processes and other use of business data takes place from a separate enterprise data warehouse (S-DWH for validated micro data), where the edited and approved data are transferred from the production database. A majority of the persons that use business data only have access to the enterprise data warehouse. All statistics produced from the integrated data warehouse are compiled based on the common unit structure and common variables.

Eurostat's Centre of Excellence on Data Warehousing[3] group has created a four-layer template, where the integrated production system can be described at a conceptual level. Every layer contains the sub-processes linked to one of the phases of the GSBPM (Generic statistical business process model). The template and the related concepts are described in more detail in the manual for statistical data warehouses, *Design and set up a statistical data warehouse*[4]. The YTY data warehouse is described in accordance with the template in the figure below that summarises the key features of Statistics Finland's integrated production system. They are described in more detail in the following sub-sections.

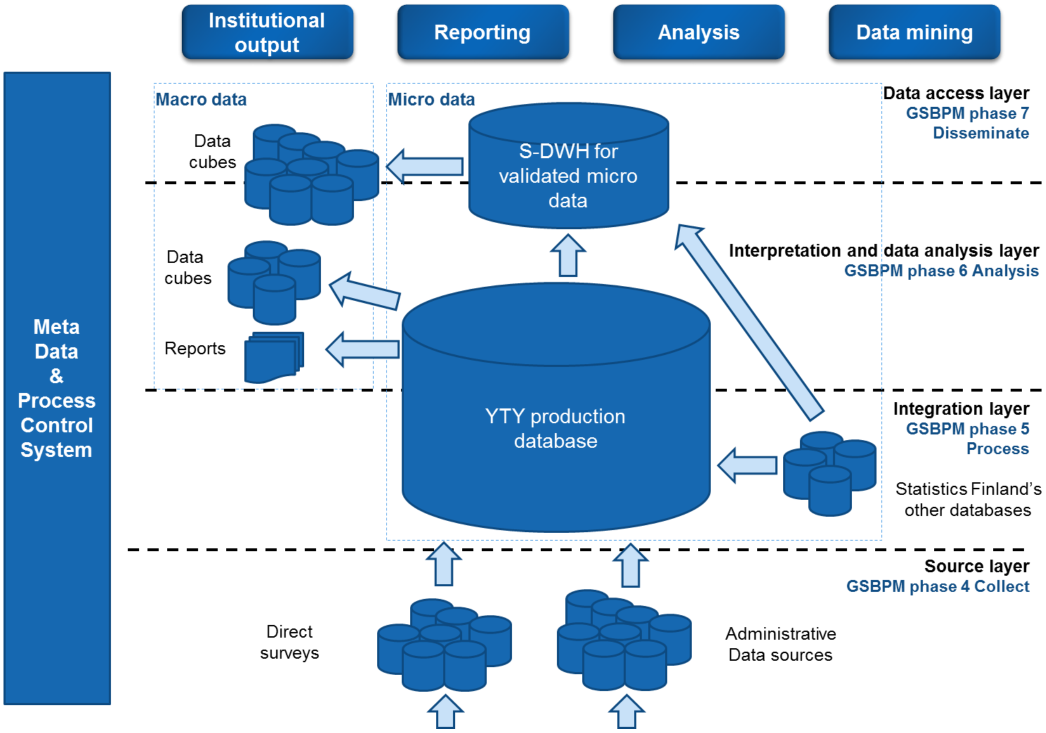

Figure 1. Integrated YTY production system.

2.1 Source layer (GSBPM phase 4 Collect)

The system has two standardised processes for reception of data, one for reception of administrative data and one for direct data collections. All received data are stored in raw data warehouses before being updated into the actual production database.

[3] https://ec.europa.eu/eurostat/cros/content/centre-excellence-data-warehousing_en

[4] https://ec.europa.eu/eurostat/cros/content/coc-dwh-s-dwh-handbook_en

2.1.1 Reception of administrative data

Sixteen different administrative data are updated into the integrated YTY production system. In terms of statistical production, the key data are data supplied by the Tax Administration. Some of the data are delivered daily, some monthly and some annually so the rate of automation of the process is high. Administrative data are delivered to Statistics Finland as text files or Statistics Finland can retrieve them as text files from the data supplier's data transmission servers. The structure, variable formats, names and permitted values of the text files are stored in Statistics Finland’s metadata system. The data are converted into SAS files based on the data in the metadata system. At the same time, it is checked that the structure of the received files is as agreed and a report is produced from the key variables. Unedited data converted into SAS files are stored in the raw data warehouse.

2.1.2 Reception of direct data collection data

There are ten direct data collections associated with the Business Register or business statistics produced from the YTY system. The direct data collections are carried out with the Xcola system built at Statistics Finland. Xcola is a browser-based data collection application with which data providers report the responses for direct data collections. The data are transferred from the Xcola system interface to the raw data warehouse of the information system using SAS DI Studio software.

2.1.3 Raw data warehouses

All source data of the integrated business data production database are stored into the raw data warehouse. The raw data warehouse contains unedited source data with which the editing process can be restarted from the beginning, if necessary. With the data in the raw data warehouse, users can also compare data that have been edited later in the process with the values of the original data. Data are stored as SAS tables in the raw data warehouse.

2.2 Integration layer (GSBPM phase 5 process)

The source data that are stored in the raw data warehouse are updated in the integrated production database. At the same time, the data are edited in accordance with the definitions of the production database. The production database of business statistics is an MS SQL Server 2012 database.

2.2.1 Integration of data into one database

The input data from different sources are integrated into the production database. Automated procedures select the data for the target variables in the production database from suitable source data and save information on what source the data derived from. Variables that are updated into the database are flagged as unverified. Data are updated with SAS DI Studio software.

2.2.2 Data editing

The variables updated into the database in phase 2.2.1 that are flagged as unverified are checked with automatic editing methods. The checked observations are marked as valid or erroneous. Erroneous observations are corrected automatically if possible or they are corrected manually by the specialists. New derived variables and estimates are calculated based on the variables updated into the database. The key variables of annual data are explored with methods of selective editing and suspicious variables are flagged to be edited manually. All of the above-mentioned functions are performed with modular software. The same editing processes are automatically performed for all data updated from different sources. The data editing processes are carried out with SAS DI studio, .NET and as SQL procedures. Observations that have been flagged as erroneous or suspicious are edited manually with a .NET application developed at Statistics Finland for this purpose.

2.3 Interpretation and data analysis layer (GSBPM phase 6 Analysis)

All data editing takes place in the integrated production database of business statistics described in Section 2.2. The separate enterprise data warehouse for validated micro data is a read only database from which users can extract edited and approved business data.

2.3.1 Transfer of data into the data warehouse

Data edited and approved in the production database are transferred to the enterprise data warehouse (S-DWH for validated micro data), where they are available for extractions by users. The transfer is carried out automatically once a day during the night. In addition, the transfer can be started by users at any other time. The transfer into the data warehouse is carried out with MS SSIS software.

2.3.2 Data analysing

Data are analysed both in the production database and in the data warehouse. OLAP cubes that automatically aggregate the data to the desired level have been built into both databases. With the help of these cubes, users can drill down into the data of suspicious aggregates, all the way to unit level. The main browser for the cubes is MS Excel.

2.4 Data access layer (GSBPM 7 Disseminate)

The data that have been edited and validated in the production database are available to users in the enterprise data warehouse. Some data are also imported to the enterprise data warehouse from Statistics Finland’s other production databases. The long-term objective is to bring all of Statistics Finland's enterprise data into this data warehouse. Users can link publication processes and other processes that use business data to the data of this data warehouse. Data can be selected both at unit and aggregate level through the OLAP cubes. In addition, users can browse the data at unit level with an application tailored for this purpose. The data warehouse always contains valid up-to-date data. In addition, the data of the data warehouse are versioned at unit level in connection with each statistical release. The enterprise data warehouse is an MS SQL Server 2012 database.

2.5 Metadata & process control system

The processes related to all steps described in Sections 2.1 to 2.4 are carried out centrally through the process management system. Users do not run any software related to editing of business data outside the process management system. All data related to the system are described in the metadata system that is utilised in the data editing processes.

2.5.1 Process management system

The process management system controls and documents all manual and automatic tasks performed in the YTY system. The process management system has been built at Statistics Finland and tailored for general use at Statistics Finland. It is comprised of two parts:

- The commercial Tibco application that defines the processes to be performed
- The Process engine X application tailored at Statistics Finland that performs the processes and saves the performance history of the processes.

Users use the Tibco application to define all processes to be performed in the YTY system. The process descriptions are made in accordance BPMN notation and the process descriptions are saved as XML files. Information on the automatic software procedures can be linked to the various process phases, these can be SAS, .NET, SQL or SSIS procedures. These process descriptions act as control data for process management. The users of the YTY system can edit the processes to be performed and their order without input from IT staff.

The process descriptions defined with Tibco are updated into the Process engine X process management system. The process phases are performed automatically with the process management system in the described order. The process phases can be manual or automatic. Automatic process phases start the procedures related to them and report on the success or failure of the procedure. Manual process phases are actions performed by users that are checked as completed in the system once finished. The process management system documents the performance history of the process such as performance times and information on who performed the task.

2.5.2 Other metadata

All data in the raw data warehouse, the production database and the data warehouse are described with the Variable Editor. The Variable Editor is an application tailored by Statistics Finland that contains data on variable names and characteristics. The data of the Variable Editor are available to the software through an interface. Some of the automatic processes in the system are directed with these metadata. The data of the Variable Editor are open to all users of the YTY system. All classifiers linked to the variables are described with their values in the Classification service from where they are available to the users and automated processes trough an interface.

Some of the metadata are at system level. The derivation and complex validation rules for the variables are defined with an application tailored to the YTY system. Users can maintain rules with a .NET interface, where rules can be edited without changes to the software code. System-related documentation is available to the users through the wiki pages of the system. Users are also informed of topical issues on the front page of the wiki, like new features of the system and possible interruptions.

2.6 System users

The production system for business statistics has over 300 users. The main users of the system are the Data Collection Department and the Business Statistics unit, as well as National Accounts. In addition, all Statistics Finland’s stakeholders that need the data of the Business Register and business statistics in their processes and persons involved in maintaining the system are users. All users have reading rights to the data of the enterprise data warehouse. Around one-half of the users work with data editing and they also have editing rights to the data in the production database. The main user of the production database is the Data Collection Department that has main responsibility for producing release-ready data for business statistics and other users of business data. The main users of the enterprise data warehouse are Business Statistics and National Accounts. They are responsible for analysing data and releasing aggregates. Business Statistics are also responsible for defining the editing rules to be carried out in the production database.

3. Development costs and efficiency benefits

The planning and implementation of the YTY system took about six years. The work began at the beginning of 2009 and was completed in December 2014. The enterprise data system was taken into use in November 2013 when it was still partially incomplete. The development work continued throughout the production cycle of the first reference year 2013 all the way until December 2014. The budgeted implementation costs for the system were around EUR two million. The actual development costs were good 20 per cent higher than budgeted.

Prior to 2009, a total of 77 staff-years were spent on the production work of the Business Register and business statistics involved. The aim of the project was to generate annual savings of 10 staff-years after the implementation of the system. During the year of implementation in 2014 this objective was not attained. As a result of the implementation of the system, the work processes of the persons involved in data editing changed, which resulted in having to learn many new things. Also, the system was implemented when it was incomplete and the development work continued during implementation. After the implementation phase and after the data of the first production years have been released, it looks like the system has achieved the efficiency objectives set for statistical production. It seems like the efficiency effect produced by the system for statistical production was realised slightly more slowly than planned.

4. Experiences of the operational data warehouse

Before the YTY system was implemented, the Business Register and Business Statistics had their own stove pipe processes and production systems that enabled independent editing and decision making. After the implementation of the integrated production system, production process has become more efficient as overlapping work phases have been eliminated. On the other hand, the work is made more complicated by that now each decision has an immediate effect on the aggregates of all statistics produced from the integrated database. The following sub-sections describe the experiences based on the first three years of operative integrated system.

4.1 Data collection and processing

The biggest benefit of the integrated system comes from the fact that all editing is carried out only once and the edits are immediately available to all business statistics. Previously, the same observations were edited several times in various systems, which was inefficient and resulted in differing editing solutions in different statistical data. In addition, it was possible that the same variables were surveyed in different data collections, which increased the response burden for enterprises.

When adopting the integrated production database, the overlapping variables were eliminated from data collections and we were able to narrow the target population of data collections. The integrated production database and tools have also improved the transparency of data processing. The requirements of a broader group of users are considered in data editing, which has improved the quality of edited data. As a result of the integrated system, the production processes of business statistics have become more uniform. The processes of various statistics follow the same logic, which has increased know-how between statistics and reduced dependency on individual persons.

Editing of data for several statistics in an integrated database also brings challenges. The solutions made in data editing should be acceptable to all statistics using business data. An observation that is insignificant to one statistics can be extremely important for the aggregates of some other statistics. The challenge is how the editing requirements of all statistics that utilise business data can be considered promptly when making editing solutions. Statistics Finland has tried to solve the problem by forming expert groups to discuss key observations in terms of business statistics and their editing. Operating in an integrated production system requires better specified and scheduled production processes from the statistics that use the system. The flexibility and response capacity of the production process in unexpected problem situations decreases from the viewpoint of an individual user.

4.2 Analysis and dissemination

The biggest benefit in terms of analysis and data dissemination derives from the fact that business data can be selected from one place with a uniform unit structure. Previously, a user that combined multiple statistical data was responsible for the integration of data and related rules. The same data were integrated multiple times and each user had to solve the problems related to integration that could derive, for example, from different unit structures in the data. The enterprise data warehouse has improved the efficiency in the production of statistics that use business data and also the coherence between such statistics. The integrated system has also increased the transparency of statistical production because analysis between statistics has become easier.

For the users of the system it has been critical that the published data are versioned at unit level. This enables analysing of published data afterwards and enables linking of extractions by other users to the data of a certain release. Versioning aims at ensuring coherence between various releases.

4.3 Ownership, decision-making and responsibilities

Responsibilities, ownership and decision-making processes related to statistical production processes changed considerably when the integrated production system was introduced. Previously, they were defined based on the stove pipe production process of each individual statistics. A single person could take part in almost all phases of the production process and an individual statistics was the owner of all phases in its production process. In the integrated system, the responsibilities and tasks have been divided by process phase. Every process must have defined ownership and decision-making, and tasks should be organised according to these responsibilities. The importance of defining the ownership of process phases has become emphasised in connection with the implementation of the integrated production system. When the system was being built, Statistics Finland carried out a reorganisation. Due to the above-mentioned reasons, the ownerships and related responsibilities were not fully defined when the system was implemented, which resulted in unnecessary confusion during the first year of implementation.

4.4 Metadata

The importance of metadata has become emphasised in connection with the transition to the integrated system as the number of data users has grown. Well-described metadata that are available to the software processes through interfaces have been essential in increasing the rate of automation in the system. Metadata have enabled us to remove control data from the programme code to be maintained by the various metadata systems. This has transferred the responsibility for maintaining the editing process from IT experts to data content experts that have the actual knowledge related to the data. The key metadata of the system are described in Section 2.5.

4.5 Technological Debt

Statistics Finland's technological infrastructure was out of date when the development work for the system began. Statistics Finland had outdated versions of key software on which the development work of the system was supposed to be based, These software were: MS Windows, MS Office, MS SQL server, .NET, SAS DI Studio and OLAP Cubes. The outdated versions that were in use for the above-mentioned software could not be used to generate solutions required by a modern information system. Therefore, Statistics Finland decided to update some of the key software to newer versions. Updating of an individual software to a newer version resulted in the software no longer communicating with other outdated software. Thus the updating of an individual software resulted in a series of software updates that had not been planned to be carried out within the project schedule. The updating caused challenges for the project schedule and a considerable amount of resources allocated to the system development work had to be used to test the updates of the software versions. As a result of the infrastructure modernisation carried out in connection with the development work, Statistics Finland now uses a newer versions of all key software. As a result of this experience, we should continue to update software to the latest versions also in the future.

4.6 Maintenance of the system

Organising of extensive system maintenance and further development is critical in order to ensure correct operations also in future. During the implementation, users provided a lot of feedback on how the system should be changed and developed further. Even after the initial implementation years we have received many development requests. Statistics Finland has reserved certain resources for further development of the system that are now allocated to development tasks as presented in the organisation in Figure 2.

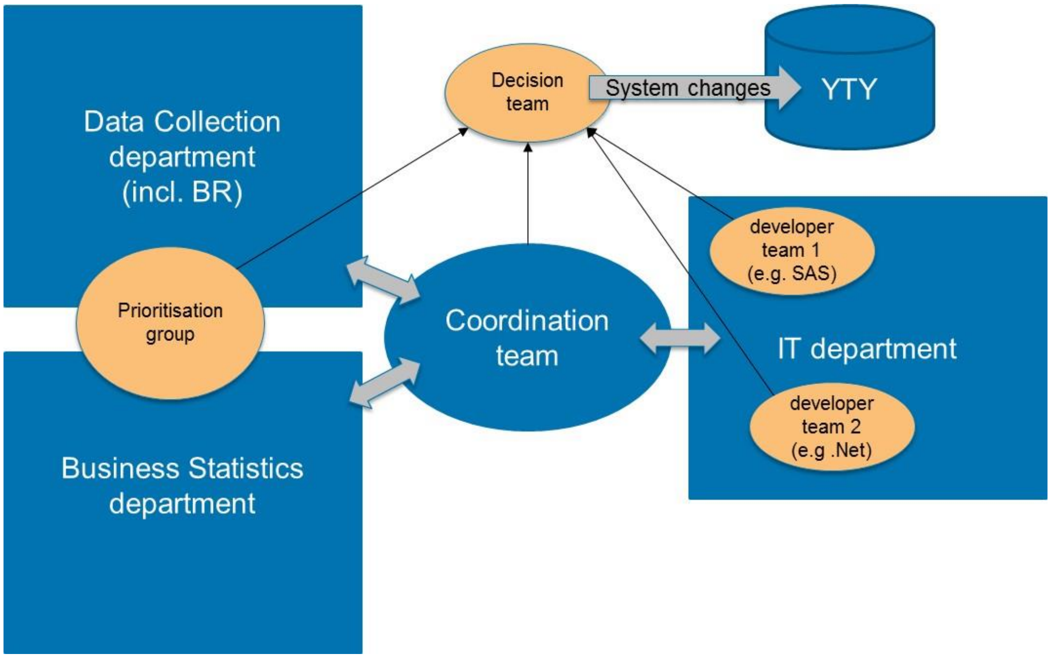
Data Collection department (incl. BR)
Decision team
System changes
YTY
Prioritisation group
developer team 1 (e.g. SAS)
Coordination team
IT department
developer team 2 (e.g .Net)
Business Statistics department

Figure 2. Organisation of the YTY system maintenance

The maintenance work of the YTY system is coordinated in the coordination team of the system that is responsible for information flow between the IT department and system users. The users of the YTY system in the Data Collection department and Business Statistics units form a prioritisation group that records all change requests and prioritises them into a backlog. The coordination group ensures that the development ideas are feasible and is responsible for more detailed definition of these change requests. The IT department has technology-specific department teams with appointed persons responsible for the maintenance tasks. The coordination team, the prioritisation group and the technology team have a representative in the decision-making team. The decision team decides how many maintenance tasks from the backlog can be carried out with the resources allocated to the maintenance of the system. These changes are available to users in new versions of the systems that are released around six times per year.

References

Saarikivi, Sami 2013: Revision Project of the Business Register and Business Statistics in Finland. Proceedings of the 2013 World Statistics Congress. http://2013.isiproceedings.org/Files/IPS023-P3-S.pdf

Centre of excellence on data warehousing. https://ec.europa.eu/eurostat/cros/content/centre-excellence-data-warehousing_en

Centre of competence on data warehousing 2014: S-DWH HANDBOOK Design and set up a statistical data warehouse. https://ec.europa.eu/eurostat/cros/content/coc-dwh-s-dwh-handbook_en