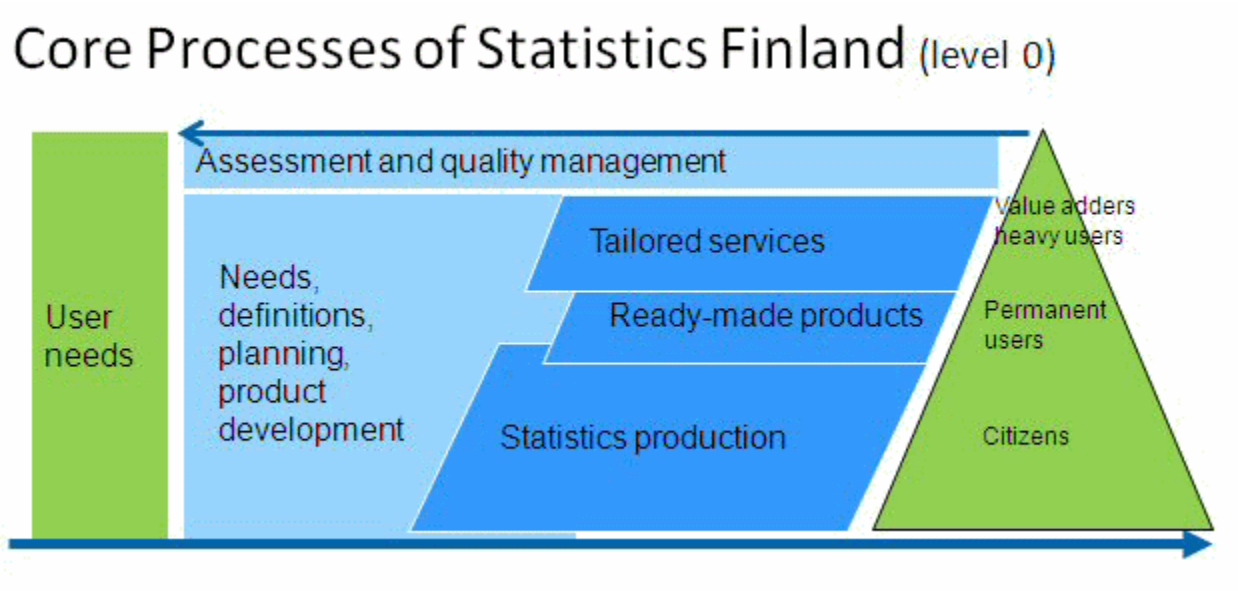
# 2. Statistical Metadata Systems and the Statistical Business Process (Finland)

Next section

## 2.1 Statistical business process model

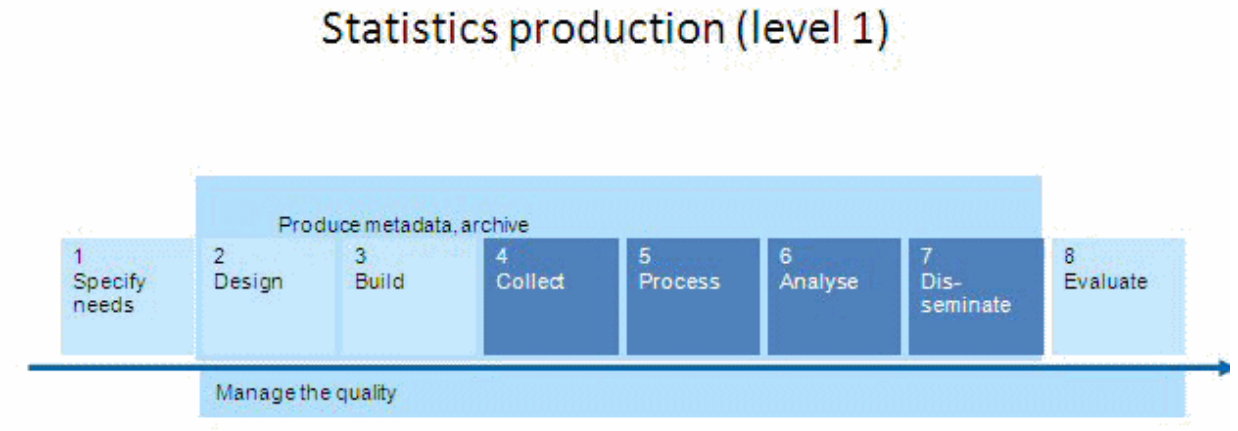
Statistics Finland has adopted a process model based on the GSBPM. We also aim to develop our model according to the lines adopted in GSBPM development. In our model however, we have come up with a solution which differs from the GSBPM in two respects: 1. To deal efficiently with the output of the business process, we have divided the end part of the business process into three different processes. These are: 1) the production of basic statistical output, 2) production of ready-made products for specific user groups, and 3) production of tailor-made services for customers. The output of the statistics production process is defined as basic publishing of statistics on the web and production of printed publications, which are intended for general use.




Of these three processes, of course, the statistics production process is the most important. The description of the two other processes aims at using the same method as the description of the statistical business process.

We feel that this solution puts focus on product development and the management of customer service. The main effects of this solution are seen in the content of the Dissemination phase, which in our model is structured in a way different from that of the GSBPM.

2. We have defined archiving as an overarching process with a close connection to metadata development and use. A corresponding solution has been made by some other countries, too.




The model was developed on the basis of our technical production model work on the one hand and on our marketing work and segmentation on

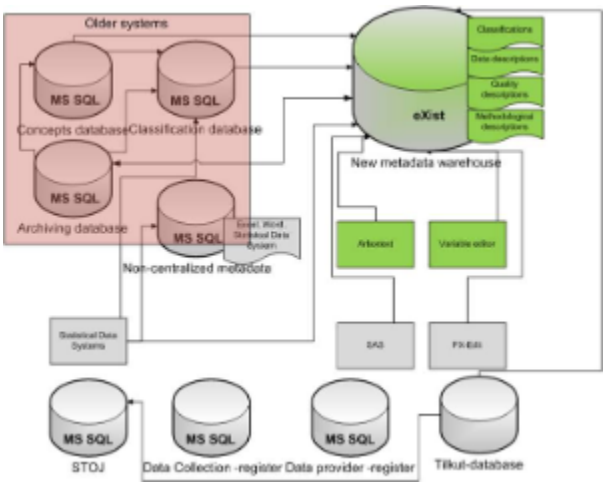
the other. In 2009 we were able to link our work to the GSBPM, which has proved to be a useful model and method worth applying.

The process map was authorised on the first level of processes in the beginning of 2009. The process description includes a description of the second level, but it is not formally authorised, because we think it would be better to leave room for flexibility if operating procedures or tools require modification.

(click on the thumbnail to view full size image)

In the near future, the work is aimed at developing a process-oriented management practice in connection with the practices of management by results. It also requires the production of a process handbook containing the most important rules for procedures and quality criteria for the various phases of the processes.

## 2.2 Current metadata system(s)




Statistics Finland's present metadata system comprises of an older part, which consists of Microsoft SQL Server databases and their PowerBuilder interfaces and of a new part currently under construction, where the eXist-XML database acts as the metadata warehouse. The new maintenance tools of metadata contents are being built. Of them, the variable editor indended for maintaining data and variable descriptions has been completed.

In practice, we thus currently operate in two different environments. Ensuring the interoperationality between them is a challenge to the development projects of the metadata system.

The currently used metadata system built in the 1990s is composed of the following parts:

- The classification database and its user interface
- The concepts database and its user interface
- The archiving database system and its user interface

The databases were originally Sybase databases that were transferred in the conversion of 2011 to the SQL Server environment.

The content of the classification database has long been utilised as SAS formats and to an extent in the statistics production processes. In the variable editor and the archiving database system, variable-specific classifications can be retrieved from the classification database and they can be added to the data description. The classifications published on the classifications pages of the stat.fi service are also produced from the classification database.

The concepts and definitions published on the stat.fi website in the concepts service are produced from the concepts database. In the variable editor and the archiving database system, concepts can be retrieved from the concepts database and they can be combined to variable descriptions.

The new metadata system elements already in use are the eXist-XML database acting as the metadata warehouse, the variable editor and the Arbortext text editor. The classifications and concepts are automatically copied from SQL Server databases to the new metadata warehouse.

The document metadata are maintained in eXist with the Arbortext text editor. Arbortext reads trilingual variable data into the tables inside the publications.

The data and variable descriptions stored in the metadata warehouse are drawn up with the variable editor. As far as possible, data maintained in other systems (classification and concepts database, the operational guidance and planning system STOJ) are used in the descriptions.

The data and variable descriptions in the metadata warehouse are utilised in trilingual tabulation in SAS and PX-Edit.

Only part of all the metadata generated at Statistics Finland are updated at the moment in the common metadata warehouses. Plenty of metadata is stored in the data systems of specific sets of statistics, in SAS and Word and Excel files, which makes them available only to the statistics concerned or even only to a certain expert. Deficient and non-uniform descriptions of metadata restrict their retrievability and usability.

## Ongoing development projects

The new metadata system aims to enhance the connections of the metadata warehouse to the statistics production process by making the metadata maintenance tools easy to use, by improving the connections of the statistical information systems to the centralised metadata warehouse, and by increasing services related to the use of metadata contents.

### *Service interface of classifications*

A whole connected to the use of classifications is initially implemented from the service-based metadata architecture (see Section 4.1). The project on the service interface of classifications defines and implements the services connected to the maintenance and use of classifications and classification conversion keys in the course of 2013.

### *Implementation of the variable editor*

In the implementation project of the variable editor, which was concluded at the end of 2012, Statistics Finland's statistical units described data sets and variables contained in them to the metadata warehouse (eXist database) with the variable editor. The Metadata Services unit trained and supported the producers of descriptions and prepared instructions together with the Information Technology and Information Services Departments. A total of 189 persons were trained during the project. In all, 123 of the sets of statistics, or slightly over 60 per cent, prepared descriptions to the metadata warehouse. At the end of the year, the metadata warehouse contained in all over 700 data descriptions. The data descriptions mostly related to the data acquisition and dissemination phase of the statistical process. The quality control of the content of data descriptions will be developed further during 2013.

After the project ended, the implementation will continue in other projects focusing on the development of the statistics production process, such as developing the reception system of administrative data.

The variable editor developer group works in connection with the projects. It deals with requests for development received from the users and decides whether they will be put into practice in the editor.

### *Renewal of archiving*

At the moment, statistical data are archived through several applications and user interfaces, which makes it difficult to manage the archiving process. The different services do not communicate with each other and no monitoring or reporting is built in them. Statistical units consider archiving a separate work phase from the production process, and therefore it is often overlooked.

The aim of renewing archiving is to define and describe the new archiving process and the data needed by it and implement the technical tools. The aim is to clarify and automate archiving of statistical data sets by combining them as an integral and non-delayed part into the statistics production process by utilising the data sets already described in the metadata warehouse. The project was started in spring 2013.

### *Quality reporting*

The project examines the relationship of the present quality reporting to the coming requirements, reviews the connection of Statistics Finland's metadata warehouse and metadata model to Eurostat's extended Metadata Standard (SIMS) and makes a plan for introducing the new quality reporting model. The aim is to perform quality reporting so that quality reports are no longer made separately for the EU, other international organisations and domestic users, but one quality report is used as far as possible in reporting. The project will start in May 2013.

## Other data systems related to metadata

TILKUT is a description database of statistics that contains basic data on statistics (name, description, topics, keywords, publication frequency and contact persons). The data from the TILKUT database are used in the stat.fi web service, the operational guidance and planning system STOJ, and the variable editor.

The operational guidance and planning system STOJ includes information on the names, publication times and contact persons of publications. The contact details of persons needed for data descriptions are retrieved from STOJ to the variable editor.

Starting from 2006, the data collection register contains data related to Statistics Finland's data collections. The system was originally built to serve metadata needs connected to direct data collections. The system was later extended to cover administrative data sets as well. In principle, the register should have all Statistics Finland's data collections described, but especially for administrative data sets, this objective has not been reached. The data are used in stat.fi's services to data providers, in the register of enterprise respondents and in Statistics Finland's planning and monitoring process. The register contains estimates of the burden caused by an individual data collection.

The register of enterprise respondents is a register intended for managing data collections to which samples, response data and respondent data are stored. The register is used to control whether a response has been received from a data provider and a rough estimate is given of the response burden.

For personal data collections, data on samples have already been collected for some time, but there is no actual register of them.

## 2.3 Costs and Benefits

### Costs

To estimate the need for human resources, the working hours spent on the variable editor project have been referred to.

At the first stage, a variable editor was designed and built for describing data and variables. A total of 381 working days were spent on the project, programming work accounting for 120 days. This was the first XML database application the main programmer had worked with, so part of the time was spent on becoming familiar with it. The project attained its goal behind schedule, and the number of working days exceeded by far the number planned. The biggest single reason for exceeding the number of allocated working days was that at the beginning of the project, it had not yet been decided which elements of the extensive metadata model were intended to be shared and obligatory for each set of statistics, nor had shared process definitions (stages of work) and terminology for the user interface yet been determined.

After the variable editor was completed, the project on its implementation was started for the years 2011 to 2012. According to the working time recording system, the realised number of staff-days in the project was 340, of which the project group used 253 staff-days (121 days for the project manager), the steering group 28 days and statistical experts 59 days (training and description preparation). The workload of statistical experts is possibly larger than the figure indicated here, because they have probably allocated some of their working time to codes of their statistical unit in addition to the project code.

The implementation of the metadata system requires use of the statistical expert resource for preparing descriptions. The amount of work is dependent on the quality of the existing descriptions. Plenty of time should be reserved for description work in connection with projects, as descriptions made according to a uniform data model will make work easier in many ways in future.

### Benefits

As far as statistics are concerned, metadata can be maintained in one system and ready for use for any set of statistics. Metadata maintained by other statistics can be applied, so overlapping work will be avoided. For example, this will make the harmonisation of concepts easier, as the definitions of any statistics can be consulted and compared in one place.

The metadata have been described systematically and appropriately. They are easy to find and made use of both for client assignments and within the organisation, e.g. for training new personnel.

From the perspective of information technology work: Centralised metadata systems serving a number of statistics production processes are likely to reduce the amount of resources needed for programming in new system projects. By utilising shared metadata systems in different statistical systems, more time can be devoted to design work, as each statistics system does not require a metadata system of its own. Shared systems ensure the availability of extra hands if needed: thanks to fewer systems, application designers will be able to devote more time, e.g. for working by two.

In order to gain optimal benefit from centralised systems, statistics specialists and programmers should be provided with information about their scope of application on a regular basis. Personnel must be kept informed about the possibilities of their application.

Development of the metadata system calls for versatile co-operation between statistics, IT and information experts. New contacts are formed in projects and much competence is shared between different expert groups. This has a fruitful effect on the activity of the organisation and on competence development.

In terms of statistics, metadata will be able to be maintained in one system and ready for use for any statistics. Metadata maintained by other statistics can be applied, so overlapping work will be avoided. For example, this will make the harmonisation of concepts easier as the definitions of any statistics can be consulted and compared in one place. The metadata have been described systematically and appropriately. They are easy to find and made use of both for client commissions and within the organisation, e.g. for training new personnel.

In terms of information technology work, centralised metadata systems serving a number of statistics production processes are likely to reduce the amount of resources needed for programming in new system projects. The use of shared metadata systems in different statistical systems contributes to time saving in design work as each statistics system does not require a metadata system of its own. Shared systems ensure the availability of extra hands if needed: thanks to fewer systems, application designers will be able to devote more time e.g. for working by two.
In order to gain optimal benefit from centralised systems, statistics specialists and programmers should be provided with information about their scope of application on a regular basis. Personnel must be kept informed about the possibilities of their application.

## 2.4 Implementation strategy

The implementation strategy is step-wise. The purpose is that once the new metadata system is ready for implementation, shifting to its application will happen in parallel with the general modification projects of every statistical data system.

Next section