

Case study: Statistics Netherlands

Case study: Statistics Netherlands Statistics Netherlands: use of GSBPM Statistics Netherlands: use of GSIM

Contact person*	.
Job title	
Email	
Telephone	

Summary*

Metadata strategy

Within the framework of the new business architecture (BA) data and metadata should be stored once and reused as much as possible. The aim is to provide for an office-wide service for the storage and the retrieval of data and metadata. This does not mean that all data is stored: only data with a certain quality meant for general use will be stored and described with the help of metadata. This is what is called steady-state data. Other metadata systems should use the central service as their source for steady-state data. Concerning the quality of metadata, there is an additional strategy. The core metadata for the central service consists of metadata that is used internationally, preferably EU-metadata. All other metadata should be formulated in terms of this core.

Current situation

In 2009 new statistical processes are implemented step by step according to the BA. A new organizational department dealing with the storage of data and metadata is established but not yet fully operational. The 'old' metadata systems are still operational and function more or less independently.

Metadata Classification

Metadata is classified according three criteria. The first criterion is when it is developed. Ex ante metadata is developed during the first three phases of the Generic Statistical Business Process Model (GSBPM): Specify needs, Design and Build. Ex post metadata is developed during each production run. These consist of the other phases of GSBPM. The second criterion is the function of the metadata. Here we distinguish four types of metadata: conceptual, process, quality and technical metadata. For instance, in short, ex ante process metadata will tell you how statistics should be produced. Ex post process metadata tells you how they were produced. The third dimension of the metadata classification is by their quality. Metadata in our pre-input and input bases are formulated according to the respondent's wording. Metadata in our other bases are formulated according the office standards. These also contain the international standards.

Metadata system(s)

The system that supports the office-wide storage and retrieval of data and metadata, the Data Service Center (DSC), is under development in a pilot phase. A first version is operational by the end of 2009. The DSC consists of two main components. The first component is the tailor-made classification server that stores and maintains classifications and code lists. The second component is based on a commercial documentation software package. This contains the metadata that is designed according to the SN Metadata model. The SN Metadata model is inspired by both the Swedish and Neuchâtel model and is meant to describe steady state data. To support a gradual development of the DSC and to guarantee the close connection with statistical processing tools, the SN Metadata model is based upon a separate metadata architecture. This metadata architecture also covers the transformation of data and metadata during statistical processing. At this moment SN is able to use the commercial software package without any tailor made software. With the help of the configuration possibilities of the software the SN Metadata model (as well as the metadata itself) can be stored and maintained. Using only the configuration mode is important because all new versions of the software can be used automatically without additional programming. It is yet unsure whether or not SN will need tailor made programming in the future; the aim is to avoid it. The BA requires to distinguish ex ante and ex post metadata. In the design phase ex ante metadata are formulated: they prescribe the statistical data required, including their required quality. During the production phase ex post metadata describe the statistical data that is realized (including their realized quality). Differences between ex ante and ex post are used to derive indicators about the quality of the statistical process and the statistical product. These indicators are meant to trigger possible future redesign phases. The DSC is able to store conceptual, process and quality metadata; the SN Metadata model however covers conceptual metadata only. Process and quality metadata are stored as free text. The first version of the DSC will contain ex post quality metadata. Ex ante quality metadata will be added in a next version. Further additional wishes are a more close relation between the two components the DSC consists of.

Costs and Benefits

Though SN has had a history of trial and error, the present pilot did not costs a lot resources. At the beginning of the pilot phase the SN Metadata model was implemented by 4 software engineers in less then one week. The tailor made classification server is a residual from earlier attempts.

Implementation strategy

The implementation strategy unfolds along multiple lines. In the first place, all new development projects should act according to the new BA and should take the DSC as a point of departure for the storage of their steady state data. In the second place, existing datasets should be added to the DSC if there is a need for reuse. In the third place all data arriving from outside the office (the so-called pre-input data) will be stored in the DSC.

IT Architecture

Metadata Management Tools

Data and metadata is stored with the help of Documentum, except for classifications and code lists: they are stored with the help of the Classification Server (a tailor made tool). Technical facilities guarantee that data are stored only if they are accompanied by suitable metadata (cf. the fourth BA principle).

Standards and formats

The metadata standards used inside the office are incorporated in the SN Metadata model. Documentum uses its own formats and standards though: through a configuration of Documentum, a translation is made between the SN Metadata model and the Documentum formats. Documentum has the ability to export metadata in XML and CSV file formats.

Version control and revisions

Documentum has an extensive version control mechanism. This is exploited to version e.g., changes in the definitions of variables through time and to version changes in the design of datasets (steady states).

Outsourcing versus in-house development

SN tries to minimise tailor made software development. Documentum is a universal tool for storing and retrieving files (called documents) in general, which makes it suitable to store and retrieve statistical data and metadata as a special use. Documentum is configured to store and retrieve statistical data according to the SN Metadata model. These configurations were implemented through external resources.

Sharing software components of tools

The SN Classification Server is tailor made and could, in principle, be used by other organisations. At the beginning of 2009 it was still in a testing phase, though.

Overview of roles and responsibilities

Apart from the general default roles that Documentum provides (such as Author and Owner) DSC distinguishes Metadata Administrator (responsible for, e.g., the maintenance of office-wide metadata standards, such as the classification of statistical topics) Metadata Designer (responsible for e.g., the correct design of data sets and for supplying correct variable definitions) User (allowed browsing the metadata base and extracting statistical data according to his need)

Metadata management team

.

Training and knowledge management

All metadata will be stored by the owners of the metadata with assistance of well- trained metadata experts.

Partnerships and cooperation

None.

Other issues

Maintaining good quality of metadata is considered a serious issue. ISO 11179 presents some guidelines and rules that are adopted by SN. It is a challenge though to formulate quality guidelines in such a way that metadata is interpreted the same way now as it will be at a later moment. Also, it is non-trivial to formulate such guidelines for metadata that is intended for (various categories) of non-experts. Acceptance of the Data Service Centre by the statistical divisions. The Data Service Centre introduces a new way of working with an initial additional workload for statisticians. This benefits the potential users of the data primarily and not so much their producers. This means that producers must be convinced that they are users (of other data) as well, so that they will see the overall benefit of spending time and money to describe their data in a user-friendly way. Additional Workload. The normal mode of operation is that metadata is produced primarily during the design phase of a statistic, of which the activities involved are usually part of a design project. For various reasons, storing and correctly describing existing data sets (SN's statistical history) with the use of the DSC is usually not organized in a project context however, which puts an additional workload on those departments that are engaged mainly in statistical production.

Lessons learned

- Small projects that deliver in short cycles;
- Use of external off-the-shelf software is possible without too much adjustments in specs;
- Keep in control of outsourced development activities;
- It is a challenge to formulate a convincing business case for metadata;
- Develop a metadata architecture in order to direct the development of metadata models that will be needed as new features for the storage, retrieval and transformation of statistical data will arise.

Links:

Attachments