WP 31
17 February 2010

ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN UNION (EUROSTAT)

ORGANISATION FOR ECONOMIC
COOPERATION AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

Work Session on Statistical Metadata (METIS)
(Geneva, Switzerland, 10-12 March 2010)

CASE STUDY – STATISTICS NETHERLANDS

Prepared by Max Booleman, Statistics Netherlands

1. INTRODUCTION

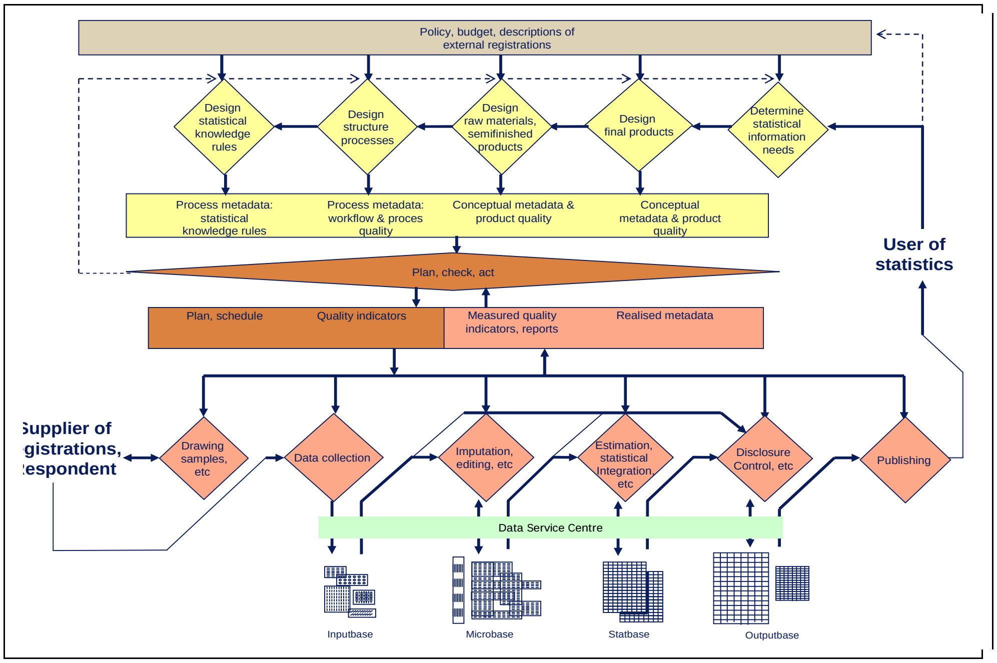
Organization Name	Statistics Netherlands
Number of staff	Approximately 2200 employees
Organization structure	Executive Board Central departments Divisionof Methodology and Quality Division of Business Statistics Division of Social and Environmental Statistics Division of Macro-economic Statistics and Dissemination Division of IT Services
Contact person (for Metadata)	Max Booleman Senior advisor/ Division of Methodology and Quality m.booleman@cbs.nl +31 70 337 4455
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.

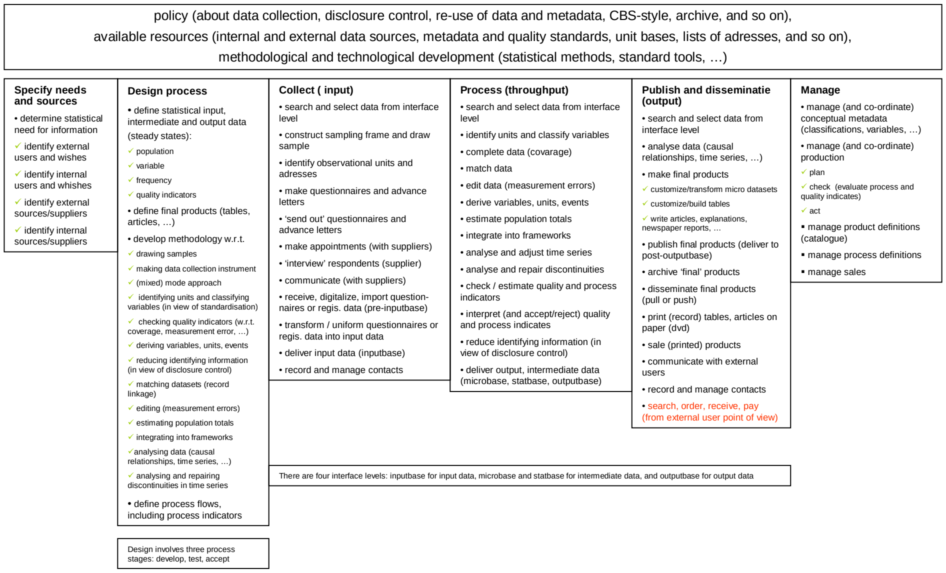
Current situation

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.

2. STATISTICAL METADATA SYSTEMS AND THE STATISTICAL BUSINESS PROCESS

2.1 Statistical business process model


Policy, budget, descriptions of external registrations
Design statistical knowledge rules
Design structure processes
Design raw materials, semifinished products
Design final products
Determine statistical information needs
Process metadata: statistical knowledge rules
Process metadata: workflow & proces quality
Conceptual metadata & product quality
Conceptual metadata & product quality
User of statistics
Plan, check, act
Plan, schedule
Quality indicators
Measured quality indicators, reports
Realised metadata
Supplier of
gistrations,
espondent
Drawing samples, etc
Data collection
Imputation, editing, etc
Estimation, statistical Integration, etc
Disclosure Control, etc
Publishing
Data Service Centre
Inputbase
Microbase
Statbase
Outputbase



policy (about data collection, disclosure control, re-use of data and metadata, CBS-style, archive, and so on),
available resources (internal and external data sources, metadata and quality standards, unit bases, lists of adresses, and so on),
methodological and technological development (statistical methods, standard tools, …)

Specify needs and sources

- determine statistical need for information
- ✓ identify external users and wishes
- ✓ identify internal users and whishes
- ✓ identify external sources/suppliers
- ✓ identify internal sources/suppliers

Design process

- define statistical input, intermediate and output data (steady states):
- ✓ population
- ✓ variable
- ✓ frequency
- ✓ quality indicators
- define final products (tables, articles, …)
- develop methodology w.r.t.
- ✓ drawing samples
- ✓ making data collection instrument
- ✓ (mixed) mode approach
- ✓ identifying units and classifying variables (in view of standardisation)
- ✓ checking quality indicators (w.r.t. coverage, measurement error, …)
- ✓ deriving variables, units, events
- ✓ reducing identifying information (in view of disclosure control)
- ✓ matching datasets (record linkage)
- ✓ editing (measurement errors)
- ✓ estimating population totals
- ✓ integrating into frameworks
- ✓analysing data (causal relationships, time series, …)
- ✓ analysing and repairing discontinuities in time series
- define process flows, including process indicators

Design involves three process stages: develop, test, accept

Collect (input)

- search and select data from interface level
- construct sampling frame and draw sample
- identify observational units and adresses
- make questionnaires and advance letters
- ‘send out’ questionnaires and advance letters
- make appointments (with suppliers)
- ‘interview’ respondents (supplier)
- communicate (with suppliers)
- receive, digitalize, import questionnaires or regis. data (pre-inputbase)
- transform / uniform questionnaires or regis. data into input data
- deliver input data (inputbase)
- record and manage contacts

Process (throughput)

- search and select data from interface level
- identify units and classify variables
- complete data (covarage)
- match data
- edit data (measurement errors)
- derive variables, units, events
- estimate population totals
- integrate into frameworks
- analyse and adjust time series
- analyse and repair discontinuities
- check / estimate quality and process indicators
- interpret (and accept/reject) quality and process indicates
- reduce identifying information (in view of disclosure control)
- deliver output, intermediate data (microbase, statbase, outputbase)

Publish and disseminatie (output)

- search and select data from interface level
- analyse data (causal relationships, time series, …)
- make final products
- ✓ customize/transform micro datasets
- ✓ customize/build tables
- ✓ write articles, explanations, newspaper reports, …
- publish final products (deliver to post-outputbase)
- archive ‘final’ products
- disseminate final products (pull or push)
- print (record) tables, articles on paper (dvd)
- sale (printed) products
- communicate with external users
- record and manage contacts
- search, order, receive, pay (from external user point of view)

Manage

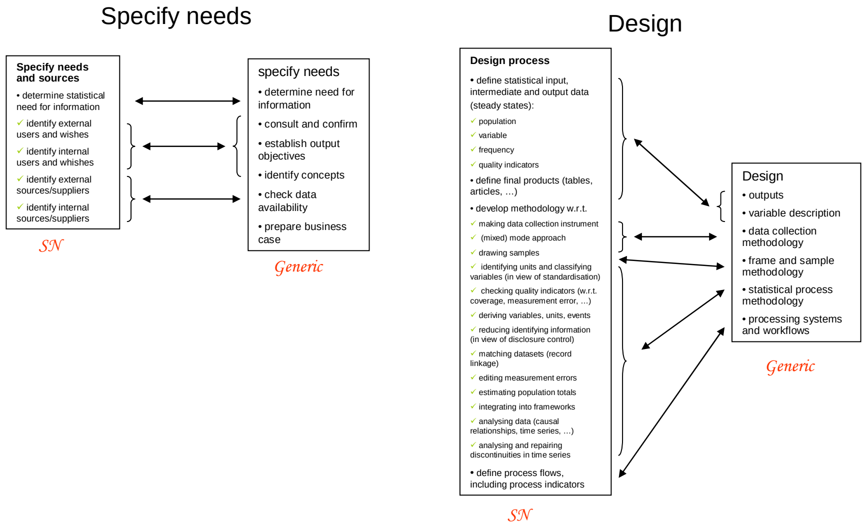
- manage (and co-ordinate) conceptual metadata (classifications, variables, …)
- manage (and co-ordinate) production
- ✓ plan
- ✓ check (evaluate process and quality indicates)
- ✓ act
- ▪ manage product definitions (catalogue)
- ▪ manage process definitions
- ▪ manage sales

There are four interface levels: inputbase for input data, microbase and statbase for intermediate data, and outputbase for output data

SN-Statistical-Business-Architecture-Model

The BA principles:

1. A strict distinction is made between the data that are actually processed and the metadata that describe the definitions, the quality and the process activities
2. No regular production takes place without controlling metadata (process model and rules)
3. During the design of the statistical process, the benefits of re-use are exploited to the maximum degree, both within and outside Statistics Netherlands
4. The metadata (of steady states) are generally accessible and are standardised as much as possible
5. The production of statistics is output-driven
6. Observational inputdata are stored as observed (in the inputbase). Publishable outputdata are stored as published (in the outputbase).
7. Steady states are traceable
8. Steady states are explicitly designed for re-use.
9. Versions of steady states are mutually related
10. A distinction is made between the statistical production process and the management of the statistical production process.


Specify needs
Specify needs and sources
• determine statistical need for information
✓ identify external users and wishes
✓ identify internal users and whishes
✓ identify external sources/suppliers
✓ identify internal sources/suppliers
SN
specify needs
• determine need for information
• consult and confirm
• establish output objectives
• identify concepts
• check data availability
• prepare business case
Generic
Design
Design process
• define statistical input, intermediate and output data (steady states):
✓ population
✓ variable
✓ frequency
✓ quality indicators
• define final products (tables, articles, …)
• develop methodology w.r.t.
✓ making data collection instrument
✓ (mixed) mode approach
✓ drawing samples
✓ identifying units and classifying variables (in view of standardisation)
✓ checking quality indicators (w.r.t. coverage, measurement error, …)
✓ deriving variables, units, events
✓ reducing identifying information (in view of disclosure control)
✓ matching datasets (record linkage)
✓ editing measurement errors
✓ estimating population totals
✓ integrating into frameworks
✓ analysing data (causal relationships, time series, …)
✓ analysing and repairing discontinuities in time series
• define process flows, including process indicators
SN
Design
• outputs
• variable description
• data collection methodology
• frame and sample methodology
• statistical process methodology
• processing systems and workflows
Generic

Mapping SN-model onto Generic model

Build

Build process

SN is striving for a situation in which a skeleton of the production process is built. For a specific production process, this skeleton is completed with specific process definitions (setups, parameters). These process definitions are developed, tested and managed by using standard tools.

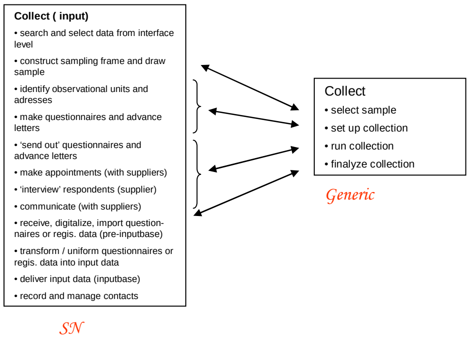
SN

Build

- data collection instrument
- process components
- configure workflows
- test production systems
- test business process
- finalize production systems

Generic

Collect



Collect (input)

- search and select data from interface level
- construct sampling frame and draw sample
- identify observational units and adresses
- make questionnaires and advance letters
- 'send out' questionnaires and advance letters
- make appointments (with suppliers)
- 'interview' respondents (supplier)
- communicate (with suppliers)
- receive, digitalize, import questionnaires or regis. data (pre-inputbase)
- transform / uniform questionnaires or regis. data into input data
- deliver input data (inputbase)
- record and manage contacts

SN

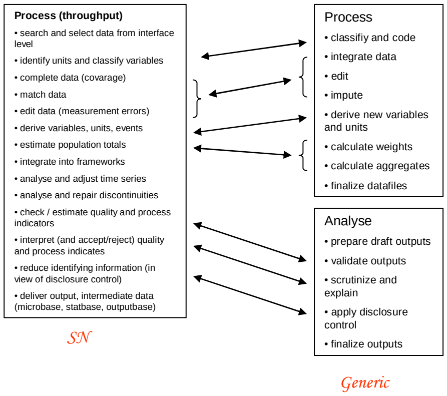
Collect

- select sample
- set up collection
- run collection
- finalyze collection

Generic

Mapping SN-model onto Generic model

Process and Analyse



Process (throughput)

- search and select data from interface level
- identify units and classify variables
- complete data (covarage)
- match data
- edit data (measurement errors)
- derive variables, units, events
- estimate population totals
- integrate into frameworks
- analyse and adjust time series
- analyse and repair discontinuities
- check / estimate quality and process indicators
- interpret (and accept/reject) quality and process indicates
- reduce identifying information (in view of disclosure control)
- deliver output, intermediate data (microbase, statbase, outputbase)

SN

Process

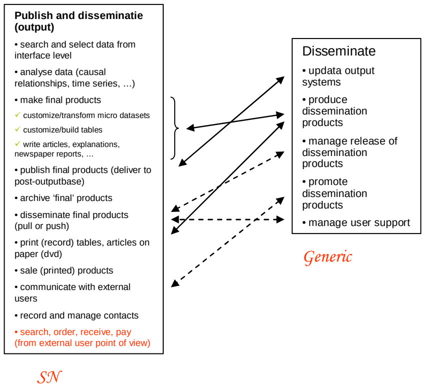
- classifiy and code
- integrate data
- edit
- impute
- derive new variables and units
- calculate weights
- calculate aggregates
- finalize datafiles

Analyse

- prepare draft outputs
- validate outputs
- scrutinize and explain
- apply disclosure control
- finalize outputs

Generic

Disseminate



Publish and disseminatie (output)

- search and select data from interface level
- analyse data (causal relationships, time series, …)
- make final products
  - ✓ customize/transform micro datasets
  - ✓ customize/build tables
  - ✓ write articles, explanations, newspaper reports, …
- publish final products (deliver to post-outputbase)
- archive 'final' products
- disseminate final products (pull or push)
- print (record) tables, articles on paper (dvd)
- sale (printed) products
- communicate with external users
- record and manage contacts
- search, order, receive, pay (from external user point of view)

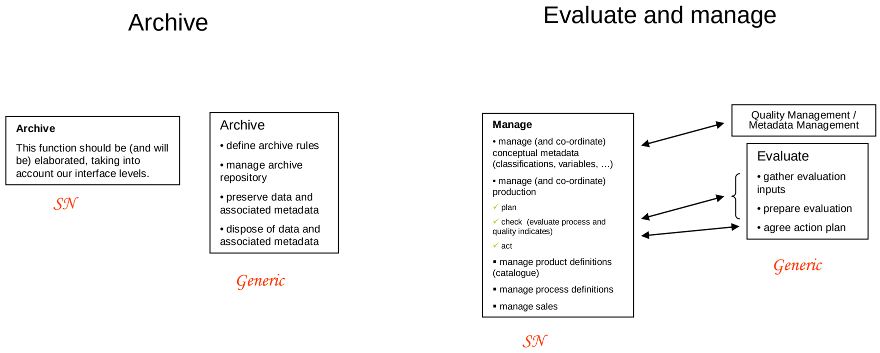
SN

Disseminate

- updata output systems
- produce dissemination products
- manage release of dissemination products
- promote dissemination products
- manage user support

Generic

Mapping SN-model onto Generic model


Archive
Archive
This function should be (and will be) elaborated, taking into account our interface levels.
SN
Archive
• define archive rules
• manage archive repository
• preserve data and associated metadata
• dispose of data and associated metadata
Generic
Evaluate and manage
Manage
• manage (and co-ordinate) conceptual metadata (classifications, variables, ...)
• manage (and co-ordinate) production
✓ plan
✓ check (evaluate process and quality indicates)
✓ act
▪ manage product definitions (catalogue)
▪ manage process definitions
▪ manage sales
SN
Quality Management / Metadata Management
Evaluate
• gather evaluation inputs
• prepare evaluation
• agree action plan
Generic

Mapping SN-model onto Generic model

2.2 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 ccomponent 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.

2.3 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.

2.4 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.

3. STATISTICAL METADATA IN EACH PHASE OF THE STATISTICAL BUSINESS PROCESS

3.1 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.

3.2 Metadata used/created at each phase

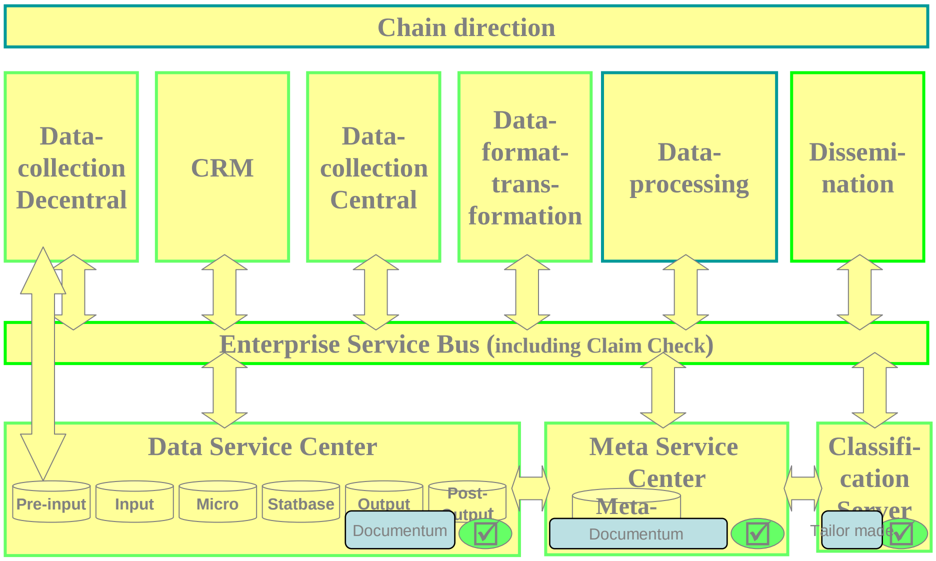
The DSC is designed to support all phases of the statistical business process. Due to its step-by-step realization, the focus at this moment is on the pre-input and input base and the output base.

3.3 Metadata relevant to other business processes

The DSC does not support non-statistical metadata.

4. SYSTEMS AND DESIGN ISSUES

4.1 IT Architecture



4.2 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).

4.3 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.

4.4 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).

4.5 Outsourcing versus in-house development

SN tries to minimise tailor made software development. Documentum is a universal tool for storing and retrievingfiles (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.

4.6 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.

5. ORGANIZATIONAL AND WORKPLACE CULTURE ISSUES

5.1 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)

5.2 Training and knowledge management

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

5.3 Partnerships and cooperation

None.

5.4 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 initialadditional 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.

6. LESSONS LEARNED

6.1

- 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.

7. ATTACHMENTS & LINKS

7.1 Talling Gelsema, Considerations for the design of the data service centre metadata model, November 2009