Standardisation of design and production of statistics:

A service oriented approach at Statistics Netherlands

Robbert Renssen and Arnout van Delden

The views expressed in this paper are those of the author(s) and do not necessarily reflect the policies of Statistics Netherlands

Discussion paper (09034)

Statistics Netherlands

The Hague/Heerlen, 2009

Explanation of symbols

.	= data not available
*	= provisional figure
x	= publication prohibited (confidential figure)
–	= nil or less than half of unit concerned
–	= (between two figures) inclusive
0 (0,0)	= less than half of unit concerned
blank	= not applicable
2007–2008	= 2007 to 2008 inclusive
2007/2008	= average of 2007 up to and including 2008
2007/'08	= crop year, financial year, school year etc. beginning in 2007 and ending in 2008
2005/'06–2007/'08	= crop year, financial year, etc. 2005/'06 to 2007/'08 inclusive

Due to rounding, some totals may not correspond with the sum of the separate figures.

Publisher
Statistics Netherlands
Henri Faasdreef 312
2492 JP The Hague

Prepress
Statistics Netherlands - Grafimedia

Cover
TelDesign, Rotterdam

Information
Telephone +31 88 570 70 70
Telefax +31 70 337 59 94
Via contact form: www.cbs.nl/information

Where to order
E-mail: verkoop@cbs.nl
Telefax +31 45 570 62 68

Internet
www.cbs.nl

ISSN: 1572-0314

6008309034 X-10

Standardisation of design and production of statistics: A service oriented approach at Statistics Netherlands

Robbert Renssen and Arnout van Delden

Summary: This paper describes the 'ideal' to-be situation of the statistical production process of Statistics Netherlands from a business perspective. Notions as 'steady state', 'unit base' and 'knowledge rule' are introduced and explained. This 'ideal' to-be situation should serve as an architectural framework for future redesigns. The paper also touches upon the latest experiences of SN to achieve this 'ideal' situation.

Keywords: architecture, business service, conceptual metadata, chain management, classification base, knowledge rule, process metadata, quality metadata, SDMX, statistical data, statistical process, unit base.

1. Introduction

Statistics Netherlands (SN) will be confronted with large challenges to decrease administrative response burden, to increase the efficiency of the statistical production process, to reduce the complexity of the landscape of IT-systems, and to control quality (see e.g. the strategic plan for the medium range, 2002-2005 and Van der Veen, 2007). Ypma and Zeelenberg (2000) describe the general approach in more detail. In order to realise these challenges simultaneously, SN has started developing architectural overviews describing the 'ideal' to-be situation.

This paper elaborates on the architectural overview that has been developed from the business perspective. The foundation of this overview is already given in Huigen (2006). It starts with a list of guiding principles that is agreed upon. Four of the most striking principles are:

- A strict distinction is made between the data that are actually processed and the metadata that describes the definitions, the quality and the process activities;
- The statistical production process is configured as a chain of process activities, controlled by means of explicitly laid down rules;
- In designing the statistical process, the benefits of re-use must be exploited to the maximum degree;
- Within the chain of process activities, we can distinguish between four fixed interface levels: the inputbase for source data, the microbase for statistical micodata, the statbase for the statistical information and the integrated macro statistics and the outputbase for publishable data.

Given the mission of SN[1] these principles reveal the chosen solution with respect to the stated challenges.

[1] The mission of SN is to compile and publish undisputed, coherent and up-to-date statistical information that is relevant for practice, policy and research.

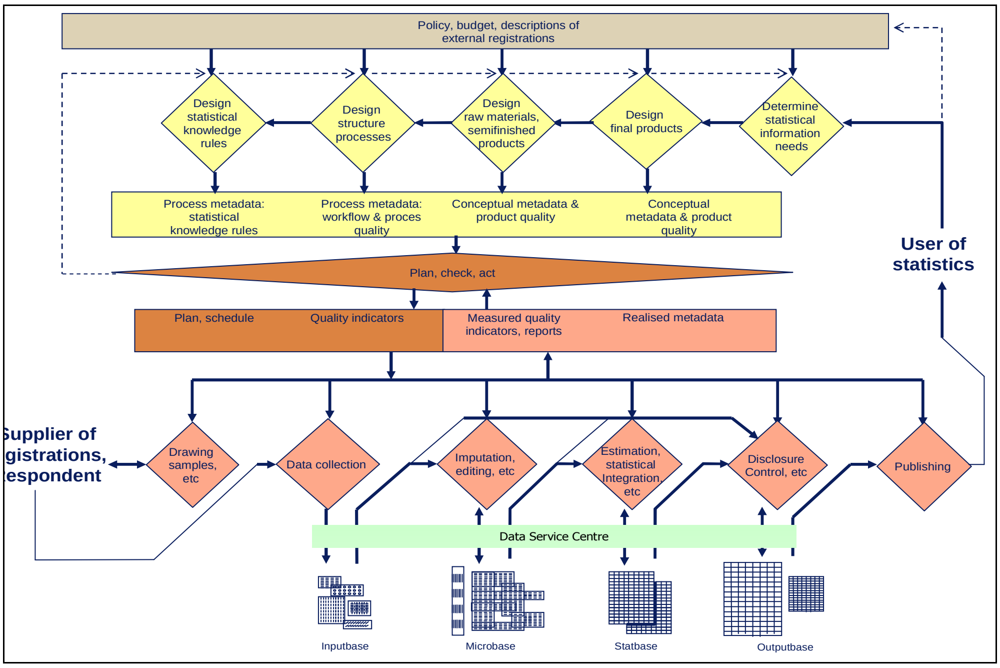
Now, the architectural overview is given in Section 2. It is further described in Section 3 and 4. Section 3 elaborates on statistical products and Section 4 on the 'ideal' statistical processes of SN. Current experiences are briefly handled in Section 5. Section 6, finally, concludes the paper.

Many ideas that are presented in this paper are not new. For example, one of the guiding principles states that statistical data that is considered as either a raw material, an intermediate product or a final product should be stored for common re-use in a so-called statistical product base. This idea has already been discussed in e.g. Willeboordse, Struijs and Renssen (2005) or Struijs (2005). Some ideas are taken over from current practice as well. For example, the concepts 'unit base' and 'classification base' that are discussed in Section 3.2 already exist and have been proven useful. New is that these ideas are explicitly formulated as principles that are agreed upon. These principles are worked out in a coherent manner, not only from a methodological point of view, but especially from a business point of view that will be followed by an IT point of view.

2. The architectural overview of SN

As the mission already states, compiling and publishing undisputed, coherent and up-to-data statistical information that is relevant for practice, policy and research, is the core business of SN. Naturally, this has to be done in an efficient manner. That is, the response burden has to decrease, the used methodology should be sound and efficient and the complexity of the landscape of IT-systems has to be simplified.

Figure 1: The architectural overview of producing statistics


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

To meet these challenges SN has developed an architectural overview, which is depicted in figure 1, see also Huigen (2006). In line with the guiding principles, there are three business domains: design (upper layer, yellow), production (lower layer, pink) and management (middle layer, red). In short, design involves designing the conceptual and quality metadata of statistical datasets as well as the (methodological) process metadata and the corresponding processes. Production involves producing statistical datasets, and managing involves planning and monitoring of the production processes as well as validating the produced statistical data. The top layer, grey, of figure 1 represents policy, the strategic goals of SN. The policy determines some of the constraints for the design stage.

Each of three business domains will be discussed in more detail. However, for the benefit of a better understanding, we will first elaborate on the nature of statistical data.

3. Statistical data

3.1 Describing real-world objects by statistical data

Statistical data can be considered as data that describe properties of real-world objects. Such an object may be a person, a household, a dwelling, or an enterprise. It may also represent a population or sub-population, like the Dutch population of male persons in 2006. These real-world objects have properties, e.g. ‘being a male person’, ‘having the age of 42 in 2007’, ‘making a profit of 20.000 Euro each year’, or ‘having an average turnover in 2007’. According to these examples one may distinguish two components of statistical data. It consists of statistical figures and of a semantic description by means of which the figures can be related to the properties of the real-world objects. This semantic description is given by the ***conceptual metadata***.

Ideally, real-world objects and their properties can be mapped during a statistical production process into a statistical dataset without errors and without time lag. Unfortunately, time lags as well as observational and processing errors are unavoidable. So, besides the conceptual metadata it is also important to accompany a statistical data-item with some quality measures, the so-called ***quality metadata***. By means of the conceptual metadata the referential meaning of a statistical dataset should be clear. Together with the quality metadata it should be possible to make statistical inferences about the real world.

The required knowledge to map real-world objects and their properties into a statistical dataset is established by the (methodological) ***process metadata***. This kind of metadata often appears in a production process as a set of parameters with respect to a general statistical method or as a set of software scripts. In this report we call both manifestations ‘***a set of knowledge rules***’.

3.2 Master data

There are two types of statistical data and metadata that deserve special attention, namely unit bases and classification bases. In short, unit bases consist of collections of identifying numbers, codes or descriptions that refer to real-world objects. Classification bases consist of collections of properties that these real-world objects may have. We distinguish simple and complex real-world objects, the latter being modelled in terms of two or more simple objects and their mutual relationships. Examples of simple real-world objects are persons, enterprises, households, dwellings, municipalities, and so on. An example of a complex real-world object is the 'triple': persons, households and their mutual relationship.

3.2.1 Unit bases and population frames

We only consider simple unit bases. In the strict sense a simple unit base can be seen as a collection of identifying numbers that refers to an umbrella population of real-world objects such as persons or enterprises that may change over time. Such an umbrella population is defined by set of properties. All real-world objects that have these properties belong to this population.

For example, at SN a simple unit base for persons is maintained. This unit base refers to an umbrella population that is defined by the property 'being a person' and further delineated by the property 'being registered into the Dutch registration of municipalities'. For reasons of disclosure control, the identifying numbers in the unit base are paired with anonymous internal numbers.

One may wonder whether a simple unit base should merely consist of identifying numbers. The answer is no. In order to determine an umbrella population for a given time period, also a limited number of specific attribute variables such as date of birth or date of death should be incorporated. By means of such variables it is possible to derive time periods in which the defining and delineating properties are valid. Then, for any given time period the umbrella population can be viewed by selecting the proper subset of identifying numbers from the unit base.

Such a subset of identifying numbers is called a population frame. The main difference between a simple unit base and a population frame is that the former can be considered as a movie of a real-world population that is changing in time, while the latter can be considered as a snap shot. Obviously, population frames that are selected from the same unit base are mutually consistent.

3.2.2 Classification bases

A classification base is a collection of classifications. We distinguish between simple classifications and hierarchical classifications. A simple classification refers to a set of properties by means of which an umbrella population of real-live objects can be divided into disjoint, i.e. mutually exclusive and exhaustive, sub-populations. When these properties are denoted by numbers, codes, or text fields describing the properties, a simple classification consists of a set of numbers, codes, or text fields.

Simple classifications can be used as the value domain of an attribute variable. They can also be used to define disjoint sub-populations.

Besides simple classifications, a classification base may also contain hierarchical classifications. A hierarchical classification consists of a sequence of nested simple classifications. Here, a simple classification A is said to be nested in B if the set of properties with respect to A can be considered a refinement of the set of properties with respect to B. For example, if ‘municipality’ is the classification that divides the Dutch population into municipalities and ‘province’ divides this population into provinces, then ‘municipality’ is nested into ‘province’. These simple classifications are related in the sense that the sub-populations defined by ‘province’ can be further divided into sub-populations defined by ‘municipality’. So, a hierarchical classification can be associated with a hierarchy of real-world sub-populations.

Together with unit bases, classification bases play in important role in designing statistical datasets in a coherent manner. On the one hand they serve as master data to define and delineate the sub-populations these datasets describe. On the other hand they are used as master data to standardise and limit the potentially large variety of properties that could be described. If two statistical datasets refer to the same real-world objects or to (virtually) the same properties, then, ideally, the same identifying numbers, codes or descriptions are used to identify these objects or to describe these properties. This enlarges the mutual comparability and coherence of statistical datasets and avoids that the end users of published data get confused.

3.3 Statistical data in steady state

During data collecting and data processing, statistical data may reach different stages of maturity. Like industrial processes we distinguish between raw materials, intermediate products and final products. Raw materials concern statistical data the way this data is collected. No statistical processing has been performed before this stage. Final products concern statistical data the way this data is published. No statistical processing will be performed after this stage. Intermediate products are processed raw materials, yet not finished for publication purposes. Like raw materials, intermediate products are deployed (or re-used) in various production processes.

Now, statistical data that is maintained either as a raw material, an intermediate product, or a final product should satisfy some requirements. These requirements concern the conceptual description and the quality of the statistical data.

- The conceptual metadata of the statistical data should be designed properly. Willeboordse, Struijs en Renssen (2006) distinguish various levels of ambition:
  - establish *well-defined concepts*
  - use *uniform language*
  - aim at *coordination* (by relating ‘nearby’ concepts)
  - aim at *harmonisation* (by reducing the number of ‘nearby’ concepts)
- The statistical dataset should meet explicitly defined quality standards.

- There is a trade-off between meeting agreed quality standards and the timeliness requirement. This conflict is solved by designing several quality versions of the 'same' dataset.
- The solution has a drawback. Too many quality versions confuse statistical users of statistics, both within SN for common re-use as well as outside SN. Therefore, each quality version should be in 'steady state' and not be replaced by a new version too quickly.

Statistical data that are released as raw materials, intermediate products or final products should satisfy these requirements. They are called steady states.

To summarize, steady states are statistical datasets that are conceptually well described, satisfy certain quality requirements and are (centrally) stored in one of the four fixed interfaces for common re-use.

4. The 'ideal' statistical process at SN

4.1 Design of steady states

Roughly, each statistical process can be characterised by a set of statistical datasets as inputs, a set of statistical datasets as outputs, and (methodological) process metadata that transforms the inputs into the outputs.

In the 'ideal' situation the outputs are defined as steady states, taking into account the needs of statistics, the budget constraints, the availability of (internal) steady states or (external) administrative data as well as the availability of statistical methods.

To meet the policy of SN with regard to coherency requirement, the design of the conceptual metadata of these outputs is supported by master data, such as the unit base and the classification base.

4.2 Management of production processes

We consider two levels of management. Process management is management within each single production process and chain management is management among a chain of production processes. Process management is at the daily, operational level. It receives orders to produce steady states from chain management and is responsible to manage the progress of the order. This involves an (operational) planning stage, a monitoring stage (checking and interpreting) and a validating (action) stage. Problems that cannot be solved at the level of process management are escalated to chain management.

Chain management is responsible for the (tactical) planning stage of orders, and to co-ordinate steady states with respect to their conceptual definitions, the quality requirements and the completion dates. Likewise this involves a (tactical) planning stage, a monitoring stage (checking and interpreting) and a validating (action) stage.

Chain management also co-ordinates possible fall-back scenarios and serves as an intermediate channel to report and inform users of steady state outputs about delays, changed quality standards or changed statistical knowledge rules. At the strategic level chain management is involved with determining the quality and timeliness standards that the output of the chain should meet and with the co-operation between the actors of the chain.

At both levels of management we distinguish between the following notions. On behalf of management, a set of *quality* and *process indicators* are defined at the design stage. On behalf of management, these indicators are *measured* at the production stage. Finally, at the management stage, these measures are confronted by *standards* that are derived from the agreed quality standards and delivery dates. Only when the quality measures do not meet the standards an action is undertaken. This action may involve a second step in a designed iterative process, a change of plan, a change of design with respect to a statistical knowledge rule, or the action may be just a change of the standard.

4.3 Production of statistical datasets

As the design processes of metadata are facilitated by the availability of centrally maintained unit bases and classification bases, the future production processes at SN will be facilitated by the use of so-called business services. In the near future SN is planning to implement three such business services, namely a Data Service Centre (DSC), a MetaData Service Centre (MDSC) and a Service for Data Collection (DC).

In order to facilitate the exchange of statistical data and metadata, SN is planning to create a DSC and an MDSC. The main services these centres offers are

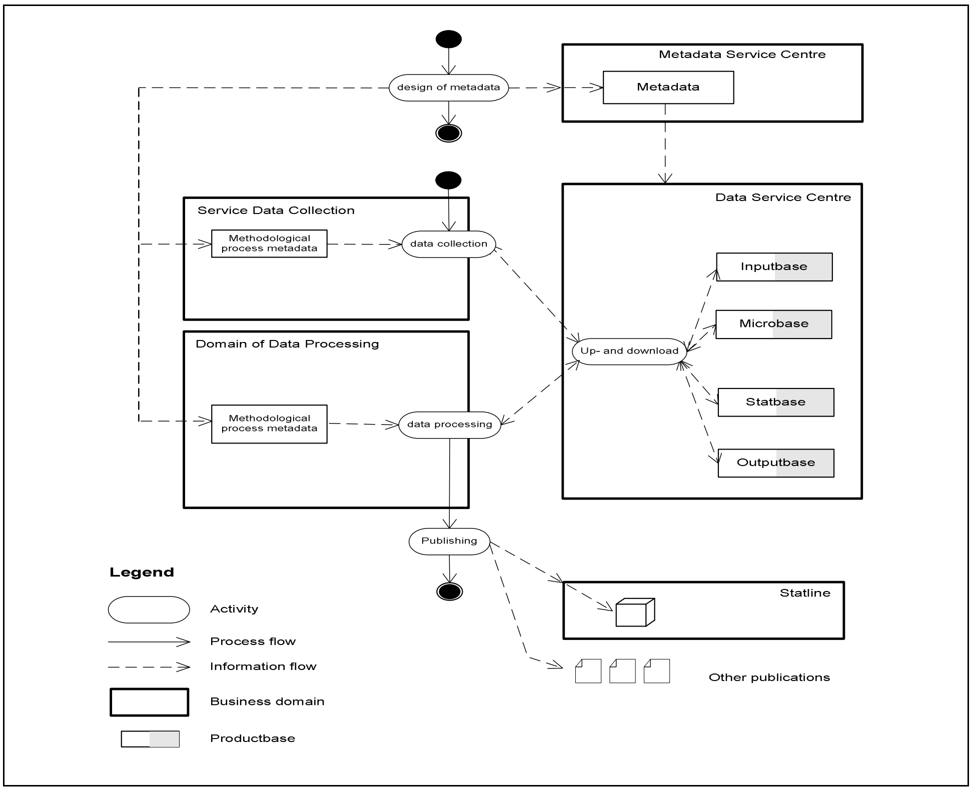
- Download, storage and upload function of the metadata,
- Download, storage and upload function of the descriptions of the steady states (in terms of the uploaded metadata)
- Download, storage and upload function of steady states.

According to one of the guiding principles there is no statistical data without metadata. This principle is enforced by DSC, since it accepts no steady states before the descriptions are loaded into the so-called catalogue. Furthermore, this catalogue accepts no metadata before the metadata is uploaded into the MDSC.

We note that there is a close resemblance between the metadata that is needed by DSC to exchange statistical data, i.e. steady states, and the metadata that is needed to exchange statistical data between national statistical organisations as described by the SDMX-model, see Sundgren, Androvitsaneas and Thygesen (2006) and Sundgren, Thygesen and Ward (2007).

Triggered by orders that are agreed upon between the owners of the production processes and DC, the service DC collects all raw data that show up in the inputbase. These data may come from primary sources as well as from secondary sources. Subsequently, these raw materials form the input of one or more production processes that process these raw materials into intermediates or final products. Once these intermediates or final products satisfy the designed quality standards, these products are stored in one of the four fixed interface levels. For publication purposes, SN distinguishes several modes (output channels), one if which is Statline.

Figuur2: Business services (near future)



In the current situation of SN there are hundreds of local applications to process data. Often, these applications can be considered as ‘black boxes’. Too few employees have knowledge with respect to the used methodology. Changing these applications, e.g. due to improved methodology, is hazardous, expensive and time consuming.

As a general solution, SN has started to implement the concept ‘rule control’: explicitly formulate the sets of knowledge rules, i.e. the methodological process metadata, and apply these rules with standard software. The guiding architectural principle behind this idea is to separate the ‘know’ from the ‘flow’, see also the second principle in the introduction of this paper. So, besides the use of business services as depicted in figure 2, the future data processing at SN will be characterized by the concept of ‘rule control’.

5. Current experiences

SN has started a Masterplan to implement both technically and organizationally the three business services. At the moment these services are not quite operational. Moreover, the scope of these services and their (technical) interfaces with the local applications are not fully clear yet.

Besides this, there is a natural reserve to use these business services. First, given the complexity of the IT-landscape it is difficult to replace only parts of the local systems and still obtain efficient processes. Second, there is a discrepancy between the short term targets of the local processes and the long term challenge of SN to increase the overall efficiency of these processes. Third, different local processes have different needs and expectations with respect to each service, which may not entirely match the scope and quality of the implemented version.

The idea of chain management is new. There is enthusiasm at SN to start using *quality* and *process indicators*. However, chain management in the sense of co-operation between different actors within a chain is for the time being hard to implement. First and most crucial, managers and employees together should identify which quality problem is first to be solved. Second, it is important to agree about the most important chain(s) of steady states that is linked to that problem. A prerequisite for that is that the most important steady states and their mutual dependencies are mapped. Third, the local process managers who ‘belong’ to a chain have to communicate and spot differences in ‘local’ interests. Fourth, so-called chain managers should be appointed to exercise supervision. Five, these chain managers should be provided with means to solve any conflicts and to take away limiting factors for achieving improvement.

Currently there are six redesigns running that are formally supported by a team of methodologists, business architects and IT-architects. In principle, these redesigns should fit into the architectural framework as described in Sections 3 and 4. These redesigns therefore start with a kind of preparation phase in which several architectural documents should be delivered and reviewed. Based on these reviews it should be clear to what extent the redesign ‘fits’ into the architectural framework.

For example, the redesign should explicitly take into account the business services MDSC, DSC and DC by explicitly designing the steady states and by formulating the so-called orders for DC. Furthermore the redesign should explicitly formulate the methodological process metadata in order to be able to choose the proper software tools, and so on. Depending on the ‘goodness of fit’, a steering group decides whether the redesign should be adapted or that SN should accept the ‘lack of fit’.

By tolerating and describing any lacks of fit it becomes possible to gradually migrate towards the ‘ideal’ situation. The sense of migrating gradually to the ‘ideal’ situation is important, because it offers the possibility to reach a consensus between short term local targets and long term overall targets.

6. Summary and future challenges

This paper has presented an architectural overview of the ‘ideal’ to-be situation of SN from the business perspective. Given the stated challenges of SN, the outline of this to-be situation is formulated by guiding principles, and further shaped in this paper. The paper has touched upon some current experiences and difficulties with implementing the business services and chain management.

SN is planning to extend the classification base with a population base and a variable base. The population base contains definitions and delineations of real-world populations that are of statistical interest. Naturally, these definitions are linked with the defining properties of the unit bases and further delineated by means of categories from the classification base. The variable base contains definitions of attribute variables. When these attribute variables are categorical, i.e. have a categorical value domain, these definitions are linked to the classification base.

Analogously to the business architecture of SN that is described in this paper, SN has developed architectural overviews from the IT-perspective. Naturally, these IT-architectures are aligned with the business architecture. We refer to e.g. the Reference Architecture Information Systems Conceptual (2007).

References

Huigen, R., Project Group Architecture (2006), ICT Masterplan; Statistics Netherlands Architecture, Business and Information model, version 1.0

Project Group Architecture (2007), Statistics Netherlands Architecture, Reference Architecture Information Systems Conceptual, version 2.0 (in Dutch).

Renssen, R.H. (2007), Producing Statistics from a Business Perspective, Statistics Netherlands.

Statistics Netherlands (2001), Statistics that Count; Strategic Plan for the Medium Range, 2002 – 2005.

Struijs, P. (2005), Improving the Quality of Statistics through the application of Process Methodology (discussion paper prepared for the advisory council on methodology and information technology), Statistics Netherlands.

Sundgren, B., Androvitsaneas, C. and Thygesen, L. (2006), Towards an SDMX User Guide: Exchange of Statistical Data and Metadata between Different Systems, National and International, Organisation for Economic Co-orporation and Development, Geneva.

Sundgren, B., Thygesen, L. and Ward D. (2007), A model for Structuring of Statistical Data and Metadata to be Shared between diverse National and International statistical Systems, second draft (work in progress to be presented at the OECD Expert group on SDMX).

Van der Veen, G. (2007), Changing Statistics Netherlands: driving forces for changing Dutch statistics (paper presented at the Seminar on the Evolution of National Statistical Systems, New York, February 2007), Statistics Netherlands.

Willeboordse, A. Struijs, P. and Renssen, R. (2005), On the role of Metadata in Achieving Better Coherence of Official Statistics, Statistics Netherlands.

Ypma, W.F.H. and Zeelenberg, C. (2007), Counting on Statistics; Statistics Netherlands' Modernization Program (paper prepared for the Seminar on Increasing the Efficiency and Productivity of Statistical Offices at the plenary session of the Conference of European Statisticians), Statistics Netherlands.