

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES
(EUROSTAT)**

**ORGANISATION FOR ECONOMIC
COOPERATION AND DEVELOPMENT
(OECD)
STATISTICS DIRECTORATE**

Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS)
(Geneva, 3-5 April 2006)

Topic (i): Metadata in a Corporate Context

THE VALUE ADDED OF A NATIONAL STATISTICAL INSTITUTE

Invited Paper

Submitted by Statistics Netherlands¹

I. CONTENT

1. The paper focuses on the necessity of good metadata and metadata models from the perspective of a process of reducing primary observations and the increasing use of external registrations. Due to this process the organisation of Statistics Netherlands changed a few years ago from a stove pipe organisation into a process oriented organisation. Increasingly, the same data source will be used by many statisticians to produce their statistics. Therefore, metadata are the corner stone of the production process.
2. Using data, it should be clear what is the content (conceptual metadata), how it is produced (process metadata) and what is the quality (quality metadata). During the production process it should be clear how it should be produced (prescribing metadata) and what has been realised so far (realization metadata).
3. The paper introduces the new business architecture of Statistics Netherlands. The role and position of metadata in the architecture will be discussed. Also organizational issues and problems related to managing conceptual metadata will be presented.
4. Statistics Netherlands is also looking for experiences within other national statistical institutes related to prescribing and realization process metadata.

II. INTRODUCTION

5. In 2000 Statistics Netherlands changed its organizational structure from a stove pipe organisation to a process oriented organisation. This organizational change was believed to better support the mission of Statistics Netherlands, which is:

“To compile and publish undisputed and coherent up-to-date statistical information that is relevant for practice, policy and research”.

¹ Prepared by Marleen Verbruggen (mrvbn@cbs.nl) and Max Booleman (mbln@cbs.nl)

6. The main reasons for this step towards a process oriented organisation were the increasing demands for statistical information and the rapid developments in the area of information and communications technology.

7. Statistics Netherlands is in the middle of important changes. The increased complexity of society and the rapid changes it is going through imply an increasing demand for reliable statistical information. The focus is shifting towards thematically presented information providing more insight how developments and sectors are related. Statistics Netherlands wants to speed up its flexible response to the changing needs for information, which requires combining data from various sources.

8. The data collection is shifting from primary to secondary collection. Developments in ICT have made most administrative data and registrations available in electronic form. This enables public and private sector organisations to compile and publish statistical information based on these administrative records. As a consequence, a National Statistical Institute (NSI) is no longer the only organisation that can produce and publish statistics. Moreover, a NSI will also increasingly be forced to use these administrative records, because this is less costly than collecting the information by surveying. Besides, there is a stringent policy to reduce the administrative burden on respondents. In the near future, therefore the value added of a NSI will not be the dissemination of statistics based on a single source, because each external register owner will be able to do so much faster and in a better way.

The value added of a National Statistical Institute will be the knowledge and skills necessary to combine sources with each other.

9. Next to this, due to privacy regulations, a NSI is sometimes the only body that is permitted to do so. Furthermore a NSI is maybe the only one who is able to present a coherent picture of the whole society.

10. In order to be able to combine many administrative sources and additional survey information, a NSI needs the appropriate methodological skills, an appropriate ICT environment and... up-to-date meta information!

11. Currently Statistics Netherlands has a few ICT systems that fit nicely in the process oriented organisation. For example, StatLine is a very sophisticated database that contains all the statistical information published by Statistics Netherlands and that can be approached by the website. There exist however also a considerable number of small ICT systems that fitted nicely in the stove pipe organisation, but are becoming a mill stone in the present organisation because they are very heterogeneous and therefore very costly to maintain. Recently, Statistics Netherlands decided to develop a business architecture with the aims to reduce the number of ICT systems, to make the remaining systems more generic and facilitate the communication between different systems. This should also have a positive effect on the costs of maintaining the technical infrastructure.

12. In chapter 2 the basic principles of the business architecture will be explained. Meta information in all its forms (conceptual metadata, process metadata and quality metadata) plays an important role in this architecture.

13. In chapter 3 we will focus on the organizational consequences of this business architecture and especially the role of metadata in this respect. Key issues are co-ordination and the attribution of tasks and roles with regard to managing metadata. This chapter will be concluded with a brief overview of present experiences with managing metadata.

14. The paper concludes with a few final remarks in chapter 4.

III. THE BUSINESS ARCHITECTURE AND THE ROLE OF METADATA

15. As introduction a few words on the concept of a business architecture. A business architecture is not a well defined concept. There are numerous private organisations that have developed very different approaches to describe a business architecture. But if one tries to describe what a business architecture does it would be more or less the following:

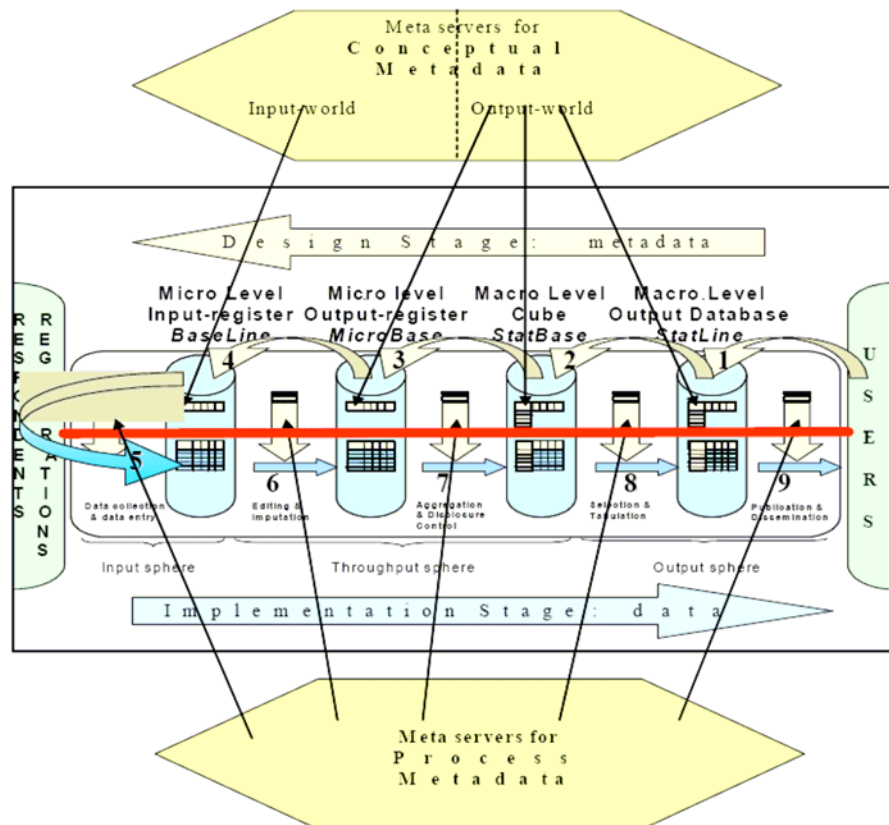
A business architecture tries to translate the strategic goals of an organisation into an optimal organizational structure including the supporting ICT systems.

16. The model that is used by statistics Netherlands is based on the Integrated Architecture Framework (IAF) and should provide the office with a decision making framework. The current view on the future set-up of the organisation shows the following principles for the architecture:

- a. Organisation: a CBS-wide, generic set-up of processes and ICT support. The ICT support must therefore be independent from the organizational structure. Research shows that primary data collection among persons and enterprises can be incorporated in one universal, common CBS model. The ICT solution has to support this.
- b. Product: support of all statistical products and a fast and flexible response to changing demands for new products. This means that ICT processes must support all statistical products, where the processes must be adjustable (relatively) fast in order to meet changes in demand.
- c. Process: preventing duplication of processes (no stove-pipes), implementing transparency by adding process data (meta!) and leaving room for exceptions.
- d. Environment: addressing the increasing diversity of suppliers by supporting a variety of channels and formats, scalability to process large quantities of input data and to add new suppliers in a simple manner, scalability for processing and production of data and preparing these data for direct integration with external sources (for instance direct integration with the basic registrations).

17. According to the current views on the business architecture, the future statistical process can be depicted as shown in figure 1. The databases in the figure, with the exception of StatLine, are “logical” databases; they do not (yet) correspond to actually existing physical databases. The figure is also simplified in the sense that it does not show the possible re-use of data and other dataflow patterns such as a feedback loops for provisional data and data integration in national accounts. However, the figure perfectly illustrates the key role played by metadata, by showing how at the design stage, ideally, user demands are translated to demand for input data (1 to 5), thereby also determining the transformations needed to get from inputs to outputs (5 to 9). For the transformations themselves, design has to be distinguished from realization of course. Although realization is meant to concur with the design, especially if the process is metadata driven, the distinction is particularly important when analysing quality aspects.

Figure 1 – Future statistical process



18. According to this model of the future statistical process, it is the metadata that has to be specified if one wants to know how inputs are processed in order to produce the statistics that are disseminated. Thus metadata (or ‘meta’) is the key to quality control. More specifically, according to the business architecture meta has the following roles:

- a. Meta describes / models the process of statistical production
- b. Meta is specified in the design phase
- c. Meta is used in the statistical production process
- d. Meta gives a univocal specification of data, process and quality (one interpretation)
- e. Meta gives a uniform specification of data, process and quality (one language)
- f. Meta is used actively in the process of statistical production
- g. Meta defines what and how and with which quality statistics are produced
- h. Meta fixes the process of data transformation in a transparent way.
- i. There are no data without meta: meta precedes the data
- j. The design of meta is the core business of a NSI.

19. In short, meta forms the corner stone in the future statistical process and the business architecture prescribes therefore the following guiding principles:

- a. A strict distinction between the implementation (production) and design (development).
- b. No regular production without meta (workflow, prescribing rules); processes are metadata driven.

In addition:

- c. There is a strict distinction between data and metadata (conceptual metadata, process metadata and quality metadata).
- d. There is a strict distinction between metadata and meta about the metadata. This last meta defines the concepts used in the metadata and can be seen as a first step towards the building of a framework of metadata.
- e. The metadata should be easily accessible and standardised (object types, variables, classifications, time dimensions, quality characteristics, process terminology).
- f. The design and maintenance of prescribing rules are important statistical activities.

20. In order to realise this broad aims and principles, SN aims for the following specific results in the next few years:

- a. The development and approval of a standardised framework of concepts of metadata, including the appropriate definitions, (the meta of metadata). This will facilitate the communication between and among users (statisticians) and software builders. A lot of preparatory work has already been done and the concept framework is almost already for approval.
- b. The design of a standardised model for conceptual metadata, based on the StatLine model (the so-called Cristal model). All statistical processes should be able to use the model and the model will describe the metadata in a univocal, uniform and systematical way.
- c. A list of demands for the ICT tools and systems that have to be bought or built to support the standardised model for conceptual metadata.
- d. Proposals to hook up several other databases to the metadata server of StatLine. This will facilitate the use of variables and classifications that are currently stored in the StatLine metadata server by other statistical processes (input and throughput).
- e. A study on the requirements on process metadata and quality metadata based on the demands of the most representative statistical processes. The study should also assess the applicability of the so called “view model” in this regard².
- f. An up-to-date and approved organizational structure for managing metadata, including a work plan.
- g. Assuring support by the most important stakeholders concerning the feasibility and advantages of a systematic description of metadata, by conducting pilots and building proofs-of-concept.

21. Most activities are strongly research oriented or ICT based, while the scope of this paper is mainly the organizational aspects of metadata. There is one important exception: managing metadata is mainly an organizational issue. The remainder of this paper will therefore focus on the concepts, plans en experiences with managing metadata.

IV. ORGANIZATIONAL ASPECTS OF METADATA

A. Introduction

22. Based on the above mentioned principles, metadata is the corner stone of the production process of statistical output. It is the most important resource of a NSI. Using data, it should be clear what is the content (conceptual metadata), how it is produced (process metadata) and what is the quality (quality metadata). During the production process it should be clear within the NSI how the data should be produced (prescribing metadata) and what has been realised so far (realization metadata).

	Prescribing metadata	Realization metadata
Conceptual metadata: <ul style="list-style-type: none">• endogenous• exogenous	X X	In principle not needed ³
Process metadata	X	X
Quality metadata	X	X

² A view is a selection of information held by any , or any combination, of the logical databases of figure 1. This translates into queries into one or more physical databases. All data processing can be seen as starting with data from one or more databases, selected by applying a view. Transformation rules are applied to this data, the result is entered to one or more databases. The result can also be represented by means of a view. Specifying views and transformations rules are an essential part of process metadata as well.

³ Only if prescribing process metadata is organising the process in such a way that the process results into the prescribing conceptual metadata.

23. In itself, managing realization metadata is not needed. These metadata should be generated automatically during the production process. What kind of realization metadata is needed to monitor, describe and present the quality of the process and the product is predefined.
24. Prescribing process and quality metadata is the result of user and process-owner demands.
25. With regard to prescribing conceptual metadata a distinction can be made between predefined external metadata (exogenous) and endogenous metadata, which can be adjusted given user wishes. Most of the time exogenous metadata equals metadata of input data. This kind of metadata should be stored and linked to (endogenous) output metadata.
26. Managing endogenous conceptual metadata is a task of a NSI and is strongly related to the statistical system. The wish to be coherent with other related concepts is sometimes in contradiction with individual internal and external user wishes. Strong internal co-ordination is needed. Statistics Netherlands distinguishes several roles to manage prescribing conceptual metadata.

B. The roles

27. For managing conceptual metadata the following roles are distinguished:
- a. **The Theme administrator.** Theme management concerns the substantive responsibility for the conceptual metadata of a thematical area (i.e. demography, employment, national accounts) and the consistency in particular. The Theme Administrator should in detail be informed on internal and external agreements concerning the conceptual metadata belonging to the theme. The Theme Administrator is responsible for compliance with national and international agreements. For a theme different Interested Parties of these metadata can be identified (see below). The Theme Administrator is also responsible for the identification of the list of Interested Parties.
 - b. **The Interested Parties.** Sub roles: Main User and remaining Interested Parties. The Interested Parties are not necessary users of the metadata; an Interested Party could use related metadata. The Main User is designated as the most important Interested Party within CBS.
 - c. **The Supplier.** Sub roles: Supplier of endogenous metadata and Supplier of metadata of secondary data sources. For all clarity: these are all internal roles, also if it concerns externally defined metadata.
 - d. **The Problem spotter.** It occurs frequently that someone observes that there is a problem related with the conceptual metadata in a specific area. A Problem Spotter reports to the Coordinator (see below).
 - e. **The Authority.** The authority is the formal owner of metadata. The Coordinator acts as delegated authority in case of unanimous recommendations of Theme Administrators.
 - f. **The Communicator.** The communicator organises the distribution and accessibility of concepts and approved metadata at Statistics Netherlands.
 - g. **The Editor of output metadata.** Disseminated metadata must meet certain standards. The Editor is responsible for this.
28. In fact with these roles management of conceptual metadata has been totally covered. Beside these roles two other roles can be distinguished to facilitate the process of co-ordination:
- a. **The Coordinator** of metadata. The Coordinator starts and coordinates the process that leads to authorised metadata. Furthermore, the Coordinator monitors the general requirements to which metadata definitions and - explanations must comply. Finally, the Coordinator is the delegated authority in case of unanimous accepted proposals.
 - b. **The Process Owner.** The production or modification of conceptual metadata from the beginning until the end can be considered as a specific process. This process has an owner, who is responsible for the progress made and the quality of the process.

C. The phases of the process

29. The process to reach authorised conceptual metadata contains different phases:
- a. **Initiation phase:** the process starts with an action aimed at approval of the metadata with a:
 - proposal by a Supplier for addition of metadata or
 - proposal by a Supplier for removal of applying metadata or
 - proposal by a Supplier for modification of applying metadata or
 - b. Signal of a problem by a Problem Spotter.
 - c. The proposal or signal will be addressed to the Coordinator, who addresses the appropriate theme and presents it to its Theme Administrator..
 - d. **Production phase:** If the Theme Administrator accepts the case, he will draw up a recommendation. In this core process the Interested Parties are involved.
 - e. **Authorisation phase:** Authorisation is no commonplace if the recommendation of the Theme Administrator is controversial. An escalation procedure is part of this process.
 - f. **Setting-up phase:** The eventual decision must be communicated within the organization, including starting date and possible accompanying measures.
30. The Coordinator has not been indicated in the process diagram. He accompanies and monitors the whole process. During the process the Coordinator ensures the communication between the Theme Administrators, Interested Parties, Suppliers, Problem Spotter and the Authority.
31. The process assumes the existence of a list of themes (including their content description) each with an overview of the Interested Parties, among which the main users. The theme list is a product of the **Theme Council**. The Theme Council exists of all Theme Administrators, a representative from the methodology department and is chaired by the Coordinator. The Theme Council meets at least once a year and establishes the list of themes (including their contents description) and actualises these. The lists of Interested Parties and Main Users by theme are maintained by the Theme Administrators. Formally there is no special role for the Theme Council in the authorisation process, but it is well conceivable that for difficult, several themes touching problems, the responsible Theme Administrator uses the Theme Council for detailed discussions. It is also possible that the Authority consults the Theme Council. Furthermore, the Theme Council decides on the annual program for the production of conceptual metadata. Due to limited resources a Theme Administrator based on this annual program could postpone or refuse new proposals by Suppliers or signal of Problem Spotters. Finally, the Theme Council is responsible for the preparation of an annual report on the progress of co-ordination of conceptual metadata.
- ### D. Experiences
32. At this moment Statistics Netherlands is running the above mentioned process without Theme Administrators. The Coordinator is acting like a Theme Administrator for all themes. One of the problems is that specialised subject matter knowledge cannot be expected from the Coordinator. It is also important to involve the subject matter departments more in the process of metadata co-ordination. A Theme Administrator should be and feel responsible for its theme. For this reason Statistics Netherlands will establish the role of Theme Administrator this year.

V. SUMMARY AND CONCLUSIONS

33. More and more external registrations become available. It is a challenge of a NSI to enrich secondary sources by combining them into a coherent and consistent set of concepts and figures. Based on this basic set a NSI is more flexible to change its output if there are new information needs. Only changing an organisation from a stove pipe into a process oriented organisation is not sufficient. ICT must also support the new processes in an efficient way. A business architecture tries to translate the strategic goals of an organisation into an optimal organizational structure including the supporting ICT systems. Multiple use of tools within a common context forces the need for a good set of appropriate metadata. This set and the knowledge of combining multiple sources is the main capital of a NSI.

34. Building a set of consistent conceptual metadata is a labour intensive job. Strong co-ordination is needed between the different thematical areas and departments. For process and quality metadata it is even not quiet clear what kind of metadata is needed and what kind of functions metadata should or could fore fill. Statistics Netherlands is looking for best practises within this field.

REFERENCES

- Altena, J.W. and A.J. Willeboordse (1997): Matrixkunde of "The Art of Cubism" (Dutch only). Statistics Netherlands, Voorburg.
- Bethlehem, J.G., J.P. Kent, A.J. Willeboordse and W. Ypma (1999): On the use of metadata in statistical processing. Third Conference on Output Databases, Canberra, March 1999.
- Keller, W.J. and A.J. Willeboordse (2001): New Methods for Statistical Processing in a New Organization Environment
- Keller, W.J. and A.J. Willeboordse and W.F. Ypma (1999): Statistical Processing in the New Millennium. Proceedings of Statistics Canada Symposium 99, Combining Data from Different Sources, Ottawa, May 1999.
- Kooiman, P., A.H. Kroese and R.H. Renssen (2000): Official Statistics: an estimation strategy for the IT-era. XIV Conference of the International Association for Statistical Computation, Utrecht, August 2000.
- Willeboordse, A..J. (2000): Towards a New Statistics Netherlands. Netherlands Official Statistics, Spring 2000, Statistics Netherlands, Voorburg.
- Willeboordse, A.J (2004): A generic framework for the structuring of conceptual meta data, Paper prepared for the meeting of the Neuchatel group, June 2004, Neuchatel