

# METADATA AT STATISTICS NETHERLANDS

*Peter Struijs*<sup>[1]</sup>  
Statistics Netherlands

## *Summary*

The use of metadata depends on the organization of the statistical production process. In the Netherlands, this process is being organized around a number of databases corresponding to the subsequent stages of statistical production. This facilitates integration within each stage of the process. The resulting needs for metadata are identified and the development of metadata tools to meet those needs discussed, with an emphasis on business statistical data. The EDI Control System and tools for secondary EDI are discussed in more detail.

Keywords: Metadata, Statistical Datawarehouses, EDI, Statistical Coordination

## **1. Background**

Official statistics in the Netherlands are centralized and traditionally much emphasis has been put on statistical coordination, especially after the second World War when the framework of national accounts became very important. One of the main achievements of coordination was the creation of a Central Business Register (the CBR), used as the coordinated frame for all business statistics. Coordination was also greatly enhanced by the development of standard classifications and definitions, although full coordination of definitions has never been realized.

In fact, these standard concepts are an early example of metadata, which can simply be defined as data on statistical data. And what is true for standard concepts is also true for other types of metadata: They can help achieving statistical coordination. However, in order to realize this, metadata need to be systematically linked to the data to which they refer, to be embedded in the statistical process and possibly to be used actively in that process. The statistical process needs to be structured correspondingly.

In the eighties the organization structure and working methods of Statistics Netherlands were still based on the traditional way of making statistics, and this environment was not favorable to embedding metadata in the statistical process. At the beginning of the nineties the decision was taken to reorganize the statistical process fundamentally. Admittedly this was not done because of a concern with metadata but rather because of changing technology, response burden considerations and user demands. The new statistical process is based on the creation of a number of statistical databases corresponding to the subsequent stages of the process. These databases are currently being implemented, which will make it possible to better integrate metadata in the statistical process.

In the next chapter the new statistical process is described in relation to the statistical databases. Chapter three discusses the metadata needs related to the statistical process and to the databases. Chapter four shows what metadata tools have been and are being developed for use in the organization as a whole, followed by a chapter on metadata for data collection as an example of the development of metadata for specific use. In the sixth chapter a couple of issues are discussed that still need to be resolved. The focus of the paper is on metadata for business statistics.

## **2. The new statistical process**

The traditional way of making statistics was to have a separate product line for each survey. A survey thus comprised the whole process from design to data dissemination, including data collection and processing. The new statistical process is being based on integration at all stages of the statistical process. Integration at the input side means that data are collected in the context of a single collection strategy for all surveys and that the results are made available for multiple purposes. One of the aims is to reduce the response burden, both in quantity and quality. Another advantage is specialization in new data collection methods and technologies, notably EDI <sup>[2]</sup>.

Integration at the output side means that statistical results from different areas are integrated and all data dissemination is carried out in the context of a single data dissemination strategy. This has many advantages, especially if clients are offered one single contact for information. Output integration has also a high potential for statistical coordination and

quality improvement. As a matter of fact, the value added of official statistics consists to a large extent of the possibility to relate data from different domains; after all, unrelated – and unrelatable – data of generally lower quality are often available to the customer from other sources.

The integration envisaged requires conceptually integrated databases at various stages of the statistical production process. Figure One summarizes the model for Statistics Netherlands as far as business data are concerned by showing these databases and their relationships. The contents of the databases is summarized in Table One. The model has not been fully implemented yet, but important progress has been made already.

The Register of Observations is meant to record all data received from primary and secondary sources. At the moment there is no such unified register yet, although copies of all observations are actually filed. The CBR registers the business population of the Netherlands, both from the input and the output side. Input units are respondent units and units as defined by administrative sources; output units are the statistical units on which statistical output is based, such as enterprise groups, kind-of-activity units and local units. The CBR also registers the links between input and output units (see Ritzen, 1995). Currently the CBR has almost full coverage of output units, but the registration of input units has so far been restricted to the units of the Chambers of Commerce and social security files. At the moment tax units are being added.

The main task of BaseLine is to record all collected information at the level of the statistical unit and to make these data centrally available to the organization. This means that BaseLine links the input variables as defined for collection purposes to the statistical units. BaseLine is already



operational, and contains, apart from CBR data, annual data on units received from the VAT register and the Corporate Tax register. By the end of 1999 monthly and quarterly VAT data will have been added; other input data will be entered later. The input variables as recorded in BaseLine are translated into output variables and recorded in the Microlab, which also contains imputed values in a number of cases where the observation is incomplete. In order to transform BaseLine data into Microlab data, a large number of operations may have to be carried out, such as validation, adjustment, derivation, imputation, and synthetic matching. The intermediate results are kept in the Throughput Databases.

*Table One: The contents of the databases (business data)*

	<i>microdata or aggregations</i>	<i>object type</i>	<i>type of count variables</i>	<i>aim</i>
<i>Register of Observations</i>	microdata	input units	input variables	full coverage of observations on input units
<i>CBR</i>	microdata	input units and statistical units	none	full coverage of all types of units and their links
<i>BaseLine</i>	microdata	statistical units	input variables	full coverage of input variables on statistical units
<i>Throughput Databases</i>	microdata	statistical units	throughput and output variables	throughput information on statistical units
<i>Microlab</i>	microdata	statistical units	output variables	all statistical information on statistical units
<i>StatBase</i>	aggregations	statistical units	output variables	all publishable statistical information
<i>StatLine</i>	aggregations	statistical units	output variables	all statistical information by area of interest

---

At present the Microlab covers statistical data on the production process only. Many Throughput Databases exist, but they are not well coordinated, mainly because of the large variety of methods used in the various surveys. The data recorded in the Microlab are not only used as input to later stages of the statistical process, but also to feed a databank used for economic research [3]. This databank is kept at Statistics Netherlands by CEREM, the Centre for Research of Economic Microdata, and may be used by researchers under very strict conditions which have to do with the confidentiality policy of Statistics Netherlands (Balk, 1998).

The aim of StatLine is to contain all statistical data published on a regular basis and to make these data accessible to external users by arranging them according to area of interest. StatLine is fully operational and covers all statistical data regularly published, comprising both business and other statistics. The data in StatLine are arranged in so-called datacubes, which are essentially multi-dimensional tables corresponding to fields of interest. Coordination of StatLine data has so far been limited; the axes of the datacubes, for instance, have not been harmonized yet. There are plans to create a database, StatBase, with all statistical data which are publishable. StatBase is to be inserted between the Microlab and StatLine. The data of StatBase have to meet conditions of confidentiality and must be sufficiently accurate.

### **3. The need for metadata**

The model of the statistical process is the basis for the metadata policy of Statistics Netherlands. Several types of metadata can be distinguished: definitional data, quality information, process information (i.e. the description of the methods by which the data are compiled such as sample schemes, questionnaires and non-response treatment), technical information, and management and control data. Sundgren (1992: 1 and 2) makes a distinction between meta-information concerning purpose and use (pragmatics), contents and meaning (semantics) and physical and technical aspects of data (syntactics). Definitional data are clearly semantic, but quality and process information comprise both semantic and pragmatic elements. Technical information is syntactic and management and control data are pragmatic.

Some metadata are needed at all stages of the statistical production process, others are needed for specific purposes. A large part of the definitional data is needed for two or more of the databases described in chapter two. This mainly concerns object types, classifications and input and output variables (plus dimensions). Quality information is database specific, but the quality of any database depends on the quality of its inputs, therefore there is a need for carry over of quality information. Process information is partly specific to the various stages of the process, but some is needed for the organization as a whole, such as a large part of the design information. And, since understanding data implies understanding how they are made, process information also needs to be carried over from one process stage to the next. Technical information is specific to the process phase concerned, but attention has to be paid where data are transferred from one process phase to the other. Detailed management and control data are also specific to the process stage concerned, but more general information is needed at higher organizational levels.

When dealing with metadata, ideally a number of conditions should be fulfilled (see also Bethlehem et al, 1998). There should be consistency and coherence not only of the data collected, processed and disseminated, but also of the metadata concerned. Ideally, metadata are integrated in the statistical process in such a way that discrepancies between ex ante (design) and ex post (realization) metadata are simply not possible. The metadata should be sufficiently complete and give, for instance, good insight in the most important aspects of quality. To the output of each database metadata should be linked that fulfill all requirements of the subsequent stage in the process. All statistical data should be reproducible, and the metadata should be sufficient to ensure this. And finally, metadata would ideally be embedded in the process in such a way that they can be used actively, i.e. can be used to control the process and generate actions.

Despite the achievements of the last couple of years in reorganizing the statistical process and in introducing new metadata tools, Statistics Netherlands is still far from the ideal situation just described. The approach chosen to get closer to the ideal is very pragmatic, there is no grand design for a completely metadata-based statistical system. There are several reasons for this. First of all, designing and implementing an overall metadata system is not only, and perhaps not primarily, a technical issue. A good system requires well coordinated concepts and methods, and experience has shown that even large portions of energy and goodwill are not enough to quickly find solutions to such questions as putting the demands of different external data users in a common framework and harmonizing methods

while keeping time series disruptions to an acceptable level. Second, there is the obvious question of resources. Apart from a huge technical effort, a large number of man-years would be needed to specify methods, enter tens of thousands of variable definitions or record all questions of all questionnaires. And last but not least, large projects tend to be difficult to realize, but huge projects seem almost never to materialize.

StatLine is a good example of the step by step approach chosen. When StatLine was designed, one of the issues was whether data that were inconsistent or whose concepts were uncoordinated should be loaded into StatLine, or whether one should first solve the imperfections [\[4\]](#). The decision was taken to load all published data at the moment of publication, irrespective of the degree of coordination. Considerations were that (1) coordination may take a very long time, (2) user demand for a database that can provide all data available was high, and (3) such a database would by itself be an important stimulus to improve coordination because it makes the problems visible for statisticians.

The incremental approach implies a need for priorities. In respect of metadata they currently include the following:

- Development of a classification server. Many classifications have been standardized for a long time, but were not available on-line until recently. StatLine needs such a tool, and a classification server is already operational (see chapter four).
- Development of a variable server. The need for such a tool is felt at many places in the organization. Different variable systems are currently operational in different parts of the organization, but they need to be integrated and their functionality improved (see chapter four).
- Registration of metainformation concerning primary and secondary EDI. For primary EDI a system with limited functionality is operational, and a project to develop metadata for secondary EDI is ongoing (see chapter five).

Other priorities, not further discussed in this paper, are the development of metadata linked to StatLine (which goes further than information on classifications and variables), and the development of quality information. An overall quality policy is being implemented at Statistics Netherlands, coordinated with Eurostat, and several quality projects are going, such as a project to develop quality instruments for data collection.

#### **4. Metadata for general use**

It is important that changes are accepted by those that have to implement them, which can be realized by offering good tools rather than issuing rules. The classification and variable servers are good examples of such tools.

##### *The classification server*

Originally the classification server was developed primarily to serve needs from the side of StatLine, but now it fulfills needs at all stages of the statistical production process (see Ypma, 1997). The most important purposes of the classification server are to store and coordinate classifications. The server is not only used for consultations, but can also be used actively in the sense that it can provide aggregation schemes and conversion keys to move from one classification to another. It is not meant as a tool to design or develop classifications, but once a classification has been entered, it ports the maintenance and allows for subsequent versions to be compiled.

Classifications are linked to object types. To each classification a classification type is assigned, for instance the type "product classifications", by which the variable of application is determined. It also has an indicator of the version (through time). If the classification is a standard classification, for any reference moment the corresponding version is the standard. Conversion (i.e. correspondence) tables refer to classifications of the same type. The most important type of correspondence table links different versions of the same classification. In some cases deviations from the standard version, the so-called variants, are entered as well. Each classification has one owner, who is the only one who can introduce new versions and who checks proposed variants.

It is interesting to see to what extent the classification server contains coordinated information. The policy applied to StatLine, to fill the database before harmonizing its content, also applies to the classification server, for similar reasons. From the outset the classifications put forward by statistical departments were accepted, but the classification server stimulates users to coordinate their classification with other classifications. At least users have accepted the uniqueness

of codes and labels, which is not a small achievement.

### *The variable server*

Operational repositories of variables exist at two places in the organization: One is built in into the EDI Control System and limited to part of the domain of business statistics (see chapter five), the other, which also contains functionality concerning object types, is linked to StatLine and has a somewhat larger scope. A project has now been started to create an organization-wide variable server incorporating the existing variable repositories with a larger coverage. In the long run the coverage is meant to include business as well as other statistics.

The variable server will not only contain definitional information including definitional relationships, but also some process and management information. It is meant to comprise, for instance, the department responsible for each variable, stakeholder departments, the variable function (markers for input, throughput and output variables) and the frames in which variables have been defined. Such a frame can, for instance, be a subsystem of national accounts, a sectorial system or a functional frame, such as environmental statistics.

As was the case with StatLine and the classification server, the policy to fill the database before harmonizing its contents is also applied to the variable server. In this way getting results is not postponed till a number of years ahead, with the risk of indefinite postponement. The price of this approach is, of course, that the results are imperfect. However, the variable server will help identifying areas where concepts need to be better coordinated, and a feasible policy would be to start making definitional relationships explicit for all variables belonging to the same domain. In fact, the variable server can be expected to become an important instrument for coordination.

## **5. Metadata for data collection**

### *The EDI Control System*

The most important existing tool containing metadata for business data collection purposes is the EDI Control System. This tool was developed when Statistics Netherlands started applying primary EDI (in the strict sense of completely electronic interchange) to businesses and represents one of the main achievements in integration of data collection. It regards the information stored in the financial bookkeeping systems of the businesses. The idea is that those businesses which are covered by primary EDI get only one electronic questionnaire, called the EDISENT questionnaire, integrating all information demand from Statistics Netherlands.

EDISENT questionnaires are created by means of the EDI Control System by using embedded metadata. This system contains repositories for type-lists, variables, questions and questionnaires, with extensive associated information concerning both definitional and control aspects. Making a questionnaire starts with defining the input variables (if not defined already), by deriving them from output variables, after which questions can be composed. It is possible to link more than one question to an input variable, because the formulation of a question may depend on the population of respondents to which the question is asked: Industries have sometimes their own "language". When formulating questions, one can cut and paste from variable and type-list class definitions and their explanatory notes. It is possible to use questions in more than one questionnaire, which can be considered an ordered set of questions. The identity of questions is very strictly defined, as is the questionnaire. Changing the reference period results in a "new" questionnaire.

Businesses receiving the EDISENT questionnaire need to establish a link between their accounting system and the EDISENT reply format. This is actually the main bottleneck in large-scale application of primary EDI. Statistics Netherlands

plies

port, but resources are limited. One way to tackle this issue is to promote the use of standard business accounting software and add statistical modules to these packages. Statistics Netherlands is trying to cooperate with the main software houses to achieve this. So far primary EDI has been applied to close to one thousand businesses belonging to a selection of 12 industries in manufacturing, public utilities, transport, and so-called business activities.

The EDI Control System was started as an experiment, and is now being integrated into the regular statistical process

(see also Keller and Bethlehem, 1998). However, its functionality is not up to large-scale survey management, and there is a need to extend the functionality with for instance the possibility to control the process of reminding.

### *Metadata for secondary EDI*

Data on statistical units collected by means of secondary EDI are stored in BaseLine, as described in chapter two. Users from statistical departments do not have direct access to BaseLine, but can send data requests to the management of BaseLine which are answered rapidly. However, some metadata on the collected data are available on-line to users in order to help them formulate their data request. These metadata are stored as intranet files (which is used as the general internal information system of Statistics Netherlands) and include an overview of variables, a description of the population and source descriptions. BaseLine itself also covers some metadata such as status codes and time variables.

So far BaseLine has had an experimental status and the data and metadata recorded are limited. Since Statistics Netherlands wants to make much more, and more systematic, use of administrative data, a project has been launched to reshape the data collection process in such a way that BaseLine can increasingly and systematically absorb administrative data (translated to the level of the statistical unit). Part of the project concerns metadata, for which the aims for the end of 1999 are the following:

- Development of the variable server as explained in chapter four. The definitions of variables from administrative sources will be entered and links with BaseLine established.
- Development of a selection tool for BaseLine. Linked to this is a tool for controlling access authorization.
- Development of technical metadata tools. The data collection process involves processing tapes received from administrative sources, using CBR data to establish the links between administrative and statistical units, compiling the variables at the level of the statistical units and putting the results in BaseLine.
- Development of tools to monitor the quality of the data collection process and products. The first steps are (1) to make an overview of quality data already available, both on the process itself and on the incoming, intermediate and resulting BaseLine products, and (2) to develop tools for measuring quality. This is a research activity which is not planned to result in computer applications in 1999, although in the longer run this is certainly the intention.

In the long run the redesign of the data collection process is meant to enable BaseLine not only to contain data collected by secondary EDI, but to include primary data as well, irrespective of the mode of collection or collecting department.

## **6. Concluding remarks**

Some general metadata issues still have to be solved, such as the question how to deal with the time dimension. At a number of places and levels the question arises when to change the definition of a concept or create a new one. This applies for instance to variables and classifications and has to do with the notion of identity [\[5\]](#). Another aspect of the time dimension is simply the question of what dates have to be recorded with the data collected. It is not difficult to think of a large number of date-type items that have some use, but it requires good judgement to apply sensible limitations. Reproducibility of data could be a guiding condition in this respect.

Another issue is how to integrate data on businesses and institutions on the one hand with data on persons and households on the other. Despite the policy to integrate all data collected at the same stage of the statistical production process instead of pursuing separate survey production lines, the separation between the two domains is remarkably persistent. Apart from the databases mentioned so far, Statistics Netherlands maintains a large database with most statistical data on persons and households (input and throughput), but this is not integrated with BaseLine or the Microlab. However, the two domains have important parts in common, such as data on persons operating their own business, or the employees of businesses. Scandinavian countries have shown how fruitful integration of the two domains can be [\[6\]](#).

It is clear that the subject of metadata is a very challenging one, but the aim to organize the statistical process in the integrated way as described in chapter two makes it even more challenging. Developing metadata requires big efforts, not only from the designers and builders, but also from those that are posed to provide and maintain the metadata, unless the metadata can be generated automatically. If a certain degree of centralization of metadata is sought, the benefits of metadata have to be made clearly visible, and not only in the

abstract. This is one of the biggest challenges for the further development of metadata.

In the history of Statistics Netherlands statistical coordination has been an increasingly important goal. Where in the past statistics were appreciated because they gave information for which there was no substitute, in the present time of information overflow their value added lies to a large extent in their coherence. In this context the development of metadata has a paradoxical effect: It tends to show, sometimes painfully explicitly, where coherence is lacking. This should not discourage the registration of metadata. On the contrary, it is one of the main means of mobilizing the forces needed to increase coordination.

## References

Bert Balk, 1998, *Establishing a center for research of economic microdata at Statistics Netherlands*. In: Proceedings of the 1998 International Symposium on Employer-Employee Matched Data, USA.

Jelke Bethlehem, Jean-Pierre Kent, Ad Willeboordse and Winfried Ypma, 1998, *On the use of metadata in statistical data processing, an overview of Statistics Netherlands' long-term strategy*, Statistics Netherlands.

Svein Gaasemyr and Peter Struijs, 1998, *The role of international standards in using register-based job files*. In: Proceedings of the 1998 International Symposium on Employer-Employee Matched Data, USA.

Wouter Keller and Jelke Bethlehem, 1998, *The impact of EDI on statistical data processing*. Meeting on the management and information technology, Geneva, February 1999.

Jean Ritzen, 1995, *Characteristics, maintenance and uses of the business register*. In: Netherlands Official Statistics, Spring 1995.

Bo Sundgren, 1992: 1, *Some properties of statistical information: pragmatics, semantics, and syntactics*, R&D report 1992: 16, Statistics Sweden.

Bo Sundgren, 1992: 2, *Statistical metainformation systems: pragmatics, semantics, and syntactics*, R&D report 1992: 17, Statistics Sweden.

Winfried Ypma, 1997, *Remarks on a classification server*, Research paper no. 9750, Statistics Netherlands.

## Footnotes:

1. The opinions expressed in this paper are the author's and are not necessarily shared by Statistics Netherlands. The author would like to thank Jean-Pierre Kent, Ben Resing, Jean Ritzen, Ad Willeboordse and Winfried Ypma for their comments.
2. The term EDI is used in this paper to denote any electronic means of data collection unless indicated differently.
3. Initially the Microlab itself was intended to fulfill this function, but later its purpose was modified into the function described in this paper; hence its name.
4. This is not to suggest a bad quality of data, but it does happen, for instance, that statistical departments produce slightly different estimates for the same variable, present the same type of data while applying different frames for aggregation or use slightly different definitions without obvious reason.
5. A similar issue is known from business registers, where decisions have to be taken on the continuation of businesses versus deletion of old businesses and entry of new ones. Inspiration for the solution may be found here.
6. Examples of national practice are given in Gaasemyr and Struijs (1998).