WP. 18
ENGLISH ONLY

**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, 9-11 February 2004)

Topic (i): Functions of metadata in statistical production

Conceptual metadata and process metadata: Key elements to improve the quality of the statistical system

Contributed Paper

Submitted by Statistics Netherlands[1]

1. Abstract

The use of the Dutch metadata model will improve the transparency of the statistical system and the dissemination policy of Statistics Netherlands. After a short introduction of the model, this paper discusses the additional possibilities for improving the quality aspects coherence and comparability of the statistical information, and the consequences for the dissemination policy. The final chapter presents the transformation process from the present data model in the Dutch dissemination tool *StatLine* to the new metadata model.

Key words: metadata model, dissemination policy, quality statistical system, conceptual metadata, process metadata

[1] Prepared by Max Booleman, e-mail address mbln@cbs.nl. The author is the statistical co-ordination programme manager at Statistics Netherlands. The views expressed in this paper are those of the author and do not necessarily reflect the policies of Statistics Netherlands.

2. Introduction

Metadata are essential in all stages of the statistical production process to identify the meaning of the data. Statisticians can only use external registrations and sources if they know exactly how concepts and classifications are defined and how they are produced. They should also know how the concepts are interrelated, so that they can link or transfer data, for instance from input concepts to output concepts.

But there is more.

A good metadata model improves the transparency of statistical information. Statistics Netherlands aims to present coherent and undisputed statistical information. With the aid of metadata it is easier to check the coherence, consistency and composition of the disseminated information.

But there is more.

One of Statistics Netherlands' aims is to become the national authority on statistical concepts. If external users have at their disposal a consistent set of concepts based on international standards and national needs, statistical information will become more effective. In the long run, external registrations will also become used to these concepts and will apply them in their own environment. If input concepts are in line with output concepts, efficiency will also increase. Statistics Netherlands is playing a very active role to stimulate the use of 'its' concepts outside the office.

3. Basics of the Dutch metadata model

This chapter presents a brief overview of the essential parts of the Dutch metadata model.

In general, a row in a statistical table represents a certain group of statistical units (a population or sub-population[2]). In the same table a column normally represents the outcome of a variable. The meaning of a cell[3], the point where the row and the column cross, can be described in terms of population, variable, period, etc. So the metadata of a cell can be determined even if it does not contain a statistical result. In the Dutch metadata model, such metadata are called *conceptual* metadata.

Conceptual metadata describe the meaning of the statistical information (borders and title of a table) and also the relationships between different kinds of concepts.

Different processes within a NSI lead to the publication of different statistical information with the same conceptual metadata. This means that within the same cell of a table, more than one figure can be published or will be available to be published. Additional metadata are needed to distinguish these different results: process metadata and quality metadata.

Process metadata describe how the statistical information is processed.

They should describe the entire process from input to output; the data sources, the assumptions, the editing procedures, applied statistical techniques etc. will provide users with valuable information on the usability of the statistical information.

Information on quality aspects of the data also gives users an idea of their usefulness. But by its nature this kind of information is different from process metadata. Unlike process metadata, quality metadata

[2] A (sub-)population can be described in terms of statistical unit and classification.

[3] Or data element according ISO-terminology

cannot be re-used. Each delivery of statistical information should be accompanied by new quality metadata. For this reason quality metadata constitute the third category of the metadata model.

Quality metadata describe the quality of the statistical information according the known quality dimensions accuracy, timeliness, punctuality, comparability, coherence.

Technical metadata are also part of the Dutch metadata model, but they are not relevant in the framework of this report.

4. Transparency of the statistical system

When the metadata model is applied, for every kind of statistical unit a table can be compiled and filled with the available statistical information. If the dimension time is included the table becomes three dimensional: a cube. There are different kinds of relationships between the cells inside the cube. The Dutch metadata model also contains relationships between the various sub-populations and the various variables. If the relationship between sub-populations (rows) and/or variables (columns) is known and fixed[4], additional statistical information can easily be calculated and presented in a cell. Within a cell, for example, annual figures based on short term indicators, on provisional annual structural statistics, or on definite annual statistics, they all have the same conceptual metadata because they all have the period(s), the population and the variable in common. Sorting the statistical information with the aid of conceptual metadata means it will be directly clear which (virtual) table cells are empty or filled with one or more statistical indicator(s).
Cubes like these are the basis of the *StatLine* presentation module. Depending on their wishes the appropriate cells and statistical information within those cells should be presented to users. Some users may be interested in statistical information based on the same process, while others want information with the highest current accuracy or the most recent release date. In this view time series could be based on the results of different processes. The following table presents cross section of the cube 'enterprises': a time series with one variable and three source-processes.

Table 1: Net turnover of enterprises by activity code

NACE activity code	2000[1)]	2001	2002	2003
…..				
industry	$STS(d,2000)$ $SBS(d,2000)$ $NA(d,2000))$	$STS(d,2001)$ $SBS(d,2001)$ $NA(p,2001)$	$STS(d,2002)$ $SBS(p,2002)$	$STS(p,2003)$
trade				
…				

1) *STS*: short-term statistics; *SBS*: structural annual statistics; *NA*: national accounts (integrated statistics); *d*: definite figures; *p*: provisional figures

At a certain moment in time the statistical information in table 1 is available and published by an NSI. Related to the other indicators[5] STS-indicators are more timely but less accurate and with less details. SBS-indicators are more accurate and detailed. Integrated figures, like National Account figures, on their turn, should present again more accurate and consistent figures about the society.

[4] Accounting systems or, more simple, definition equations are examples of relations between subpopulations and/or variables.

[5] Sometimes even more indicators, like business cycle tendency indicators or quarterly (national) accounts, are available.

Conversely, the aims of annual SBS statistics and integrated statistics (NA) are not only to increase accuracy and the level of detail, but also to increase the accuracy of the forthcoming short-term statistics or ‘now casts’.

If a user wants to have a time series, the NSI has to decide which series should be presented. It could be a series based on the STS up to 2003, SBS up to 2002 or based on the NA up to 2001. The choice should be thought out very carefully, depending also on the user’s wishes and needs.

But there is also a fourth possibility. Within a consistent statistical system STS(2003) predicts SBS(2003), and the growth rate STS(2003)/STS(2002) predicts SBS(2003)/SBS(2002). The same holds for the indicators of National Accounts. Depending on the model and the assumptions, a new time series could be constructed:

Table 2: Net turnover of enterprises by activity code

NACE activity code	2000[1)]	2001	2002	2003
…..				
industry	$NA(d,2000)$	$NA(p,2001)$	$NA(\hat{p},2002) =$ $NA(p,2001)\times\frac{SBS(p,2002)}{SBS(d,2001)}$	$NA(\hat{p},2002)\times\frac{STS(p,2003)}{STS(d,2002)}$
trade				
…				

In the present situation most NSI’s do not produce such tables. Usually a table contains only information derived from the same survey or the same kind of process. This means inconsistencies between the different indicators are concealed and therefore there is hardly any need to compile consistent estimates.

Of course this can also be achieved within the present system, but there is no trigger to treat statistical information as one interrelated system and there is no transparent statistical system to identify discrepancies. Given the present situation it is more difficult to harmonise concepts, populations and methods to disseminate consistent results.

NSI’s work with too many independent cubes. Instead of a cube for each survey process, cubes should be combined into one larger cube for each type of statistical unit.

The Dutch metadata model makes it possible for NSI’s to publish as they have done up to now, but it provides the extra possibility of a transparent view of the contents and consistency of the statistical system as a whole. The model gives NSI’s a systematic view of the statistical information, the filling of the cells and invites them to disseminate consistent estimates irrespective of underlying production processes.

For users of statistics it means that statistical information will be published more coherently and furthermore that, depending on their wishes, they can compile tables on the fly.

At a higher level even the cubes are not independent from each other. For example, household income (cube ‘households’) is related to salaries paid by enterprises (cube ‘enterprises’).

5. Improving the statistical system

The Dutch metadata model renders the statistical system much more transparent. It makes it easier to trace inconsistencies in nomenclature[6] and statistical information but, as present practice shows, extremely difficult to solve.

The use of unnecessary different sub-divisions of groups of statistical units can be reduced. The same is true for various comparable concepts. If NSI's are able to limit the number of different (similar) concepts, the available statistical information may increase for certain sub-divisions. The following tables explain the strategy of Statistics Netherlands to improve the usability of statistical information.

Table 3: A population by two different kinds of subdivisions and two different variables

Size classes	Variable X	Variable close to X
0-5	A	
6-10	B	
11-20	C	
21 and more	D	
0-15		E
16-50		F
51 and more		G

subdivision 1

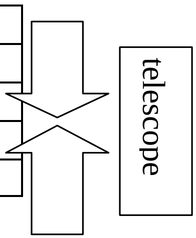
subdivision 2

If the second classification (sub-division) does not serve an explicit user need, an NSI should opt for the following table by publishing 'Variable close to X' according subdivision 1.

[6] The same name for two different concepts, or conversely two (or more) names for the same concept.

Table 4: A population by one kind of subdivision and two different variables

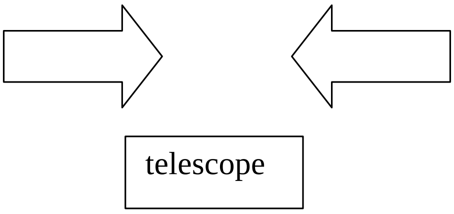
Size classes	Variable X	Variable close to X
0-5	A	H
6-10	B	I
11-20	C	J
21 and more	D	K



However, if ‘Variable close to X’ is not explicitly needed for certain users and if it is possible to transform this variable into ‘Variable X’, then an NSI should choose the following table instead of table 3.

Table 5: A population by two different kinds of subdivisions and one variable.

Size classes	Variable X
0-5	A
6-10	B
11-20	C
21 and more	D
0-15	L
16-50	M
51 and more	N



To be complete, of course an NSI should prefer the following table to table 3.

Table 6: A population by two different kinds of subdivisions and two different variables

Size classes	Variable X	Variable close to X
0-5	A	H
6-10	B	I
11-20	C	J
21 and more	D	K
0-15	L	E
16-50	M	F
51 and more	N	G

Conclusion:

It is the task of an NSI to minimise the number of empty cells within a cube to improve coherence and comparability!

6. The metadata model applied: *StatLine* 2004

Step by step the Dutch metadata model will be implemented in the dissemination tool *StatLine*. By mid 2004, the present contents of *StatLine* will be transformed to a new data model based on cubelets and the underlying *Cristal* model. A cubelet equals one column of an existing table in *StatLine*. All these cubelets are in fact very small cubes describing only one period for one variable and for a certain population which can be divided into sub-populations (the rows of the table). This stage can be done automatically. The metadata will be stored separately from the indicators.

In the second stage of the transformation process, the various cubelets will be linked to each other to become real cubes. This will be a very labour-intensive process and may take several years as the various populations and variables will have to be identified as being equal (or not).

Meanwhile new statistical information will be added in accordance with the new data structure. This means the figures can be added, but the corresponding metadata have to point to available metadata inside *StatLine*. If new metadata are needed, this should be indicated by the producing department, and a special division, responsible for the control of metadata, should add the new metadata beforehand. To increase the international comparability the international standards are leading.

The strategy of the purpose-made metadata management division is that in principle, national concepts and classifications are based on international standards.

References

Max Booleman et al., *Attention to quality within Statistics Netherlands: quantifiable quality characteristics,* Statistics Netherlands, 1999.

Jean Pierre Kent et al., *On the use of metadata in statistical data processing*, Statistics Netherlands, 2000.

LEG on Quality, *Final report*, Stockholm, 2001.

Erik van Bracht, *Cristal White paper*, Statistics Netherlands 2002.

Erik van Bracht, *CRISTAL, a Model for the Description of Statistics*, Statistics Netherlands 2002.

Neuchatel group, Neuchatel terminology Classifications, 2002.

Neuchatel group, Neuchatel terminology Variables, 2003.