What metainformation should accompany statistical macrodata?

Bo Sundgren

SCB

R&D Report
Statistics Sweden
Research - Methods - Development
1991:9

INLEDNING

TILL

R & D report : research, methods, development / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1988-2004. – Nr. 1988:1-2004:2.
Häri ingår Abstracts : sammanfattningar av metodrapporter från SCB med egen numrering.

Föregångare:

Metodinformation : preliminär rapport från Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån. – 1984-1986. – Nr 1984:1-1986:8.

U/ADB / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1986-1987. – Nr E24-E26

R & D report : research, methods, development, U/STM / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1987. – Nr 29-41.

Efterföljare:

Research and development : methodology reports from Statistics Sweden. – Stockholm : Statistiska centralbyrån. – 2006-. – Nr 2006:1-.

R & D Report 1991:9. What metainformation should accompany statistical macrodata? / Bo Sundgren.
Digitaliserad av Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-1991-X101OP9109

What metainformation should accompany statistical macrodata?

Bo Sundgren

SCB

R&D Report
Statistics Sweden
Research - Methods - Development
1991:9

Från trycket Juni 1991
Producent Statistiska centralbyrån, utvecklingsavdelningen
Ansvarig utgivare Åke Lönnqvist
Förfrågningar Bo Sundgren, 08-783 41 48

ISSN 0283-8680
Printed in Sweden

WHAT METAINFORMATION SHOULD ACCOMPANY STATISTICAL MACRODATA?

Professor Bo Sundgren
Statistics Sweden
Research & Development
S-115 81 STOCKHOLM

1991-05-30

0 The purpose of this paper

The purpose of this paper is to serve as an introduction to a discussion at the June 1991 Meeting of *Working Party 9* of the OECD *Industrial Committee*. The topic of the discussion is: *Standards for Metadata in International Databases*.

The paper is of a methodological character. It concerns the contents of the meta-information rather than EDP technical considerations. The reference list contains some papers (1, 2, 3, 5) describing the practical organization and usage of metadata in the Swedish AXIS system for handling statistical databases in a combined mainframe/PC environment.

1 Metainformation about statistical macrodata: concepts and requirements

1.1 Statistics production: basic concepts and logical structure

Definition of "statistical macrodata"

From a contents-oriented point of view **statistical macrodata** consists of

- **estimated values** of
  - **characteristics**, or **parameters**, of interest of
  - **collectives of objects** of interest, also called **populations** and **domains of interest** (where the latter are usually supposed to be subsets of the former).

End of definition

The macrodata are the result of an **aggregation process** based upon **statistical microdata**, which consists of

- **measured values** of **variables** of **individual objects.**

The statistical microdata, resulting from observations and measurements, are typically subject to a **preparation process**, consisting of coding, editing, and data entry steps, before they are submitted to the aggregation process for producing the estimates of the macrolevel characteristics.

An aggregation process transforms microdata into macrodata. Typical aggregation processes are frequency counting and summation. In the case of sample surveys, weights related to sample inclusion probabilities are used in the aggregations. Non-response, and other events and condition causing distortions in the measurement process may also force the aggregation process to be more complicated than a mere counting and summarizing operation.

Since the purpose of statistical aggregation is estimation, the micro/macro transformation process is also labeled as an **estimation process.**

1.2 Main purposes of metainformation about statistical macrodata

Metainformation is usually defined as

- information about information, and
- information about data.

The physical representation of metainformation is called **metadata.**

From a practical point of view, metainformation could be regarded as a set of **descriptions of information and data.** Here we are interested in analyzing

- Which descriptions of macrodata in a statistical database would be needed, in order to help potential users of the database to judge the usefulness of the macrodata for different purposes?

In section 1.1 above we established a logical structure of statistics production. From this structure we may identify three different, but related descriptions that need to be part of macrodata metainformation:

(1) *Firstly,* we need to describe which parameters, of which populations and domains of interest, that the statistical macrodata are intended to be estimates of.

(2) *Secondly,* we must describe

  (a) which variables, for which microlevel objects, were intended to be measured (observed) as a basis for the estimation of the macrolevel parameters; and

  (b) the "ideal" estimation function for deriving the parameter estimates

from the variable values.

(3) *Thirdly,* we should describe how the observations and measurements were actually carried out and processed, which deviations from the plan that took place, which distortions that occurred, and how this all affected the estimation process and the resulting quality of the macrodata.

The first two parts of the description are essential in order to communicate to a user of statistical macrodata an understanding of the **intended meaning** of the data. The third part of the description is essential in order to facilitate for the users to judge the **errors and uncertainties** in the actually available data in comparison with what was ideally intended.

If we could observe (measure) all individual objects in the population in accordance with the "ideal" definitions referred to in (1) and (2), and if we could assume that all these observations (measurements) were "complete" and "correct" in all respects (cf (3) above), then the estimations could be done with certainty, that is, they would not really be "estimations" in the natural sense of this word, but rather exact calculations. However, in practice both assumptions just mentioned are usually violated.

In connection with a database containing statistical macrodata from many different places (for example, many different countries) and for several time periods, we could formulate a fourth requirement on the metainformation accompanying the macrodata:

(4) *Fourthly,* we should describe the **comparability** between data in time and space, that is, in the case of OECD data, between countries as well as between time periods.

As for the "space" dimension, the comparability could probably best be described indirectly, by giving information about how the statistical macrodata from every specific country relates to relevant international standards and recommendations.

2 The logical structure of statistical macrodata

Following the definition of "statistical macrodata" in section 1.1 above, an overview of the (intended) macrodata contents of a statistical database could be given in terms of a simple **list of parameters of collectives of objects**. More precisely, the list would in the first place name the parameters of collectives of objects that were *intended to be estimated* on the basis of certain measurements of microlevel variables of microlevel objects. In order to fulfil the metainformation requirements indicated in section 1.2, the list would have to be supplemented with more detailed information about definitions of the intended parameters and object collectives, specifications of the intended measurements and estimation functions, and descriptions of the *actually performed* measurements and estimation procedures, together with some indications of the discrepancies between what was ideally intended and what was actually achieved, and the possible effects of these discrepancies on the resulting quality and usefulness of the macrodata actually stored in the database.

A list of the type briefly described above is sometimes referred to as a **tabulation**

plan. This very term indicates that many users of statistics think of statistical macrodata in terms of **tables** rather than in terms of more abstract lists of characteristics. This could sometimes be regretted, but nevertheless it has to be considered when discussing the logical structure of statistical macrodata.

2.1 Statistical tables: physical and logical structure

For the man in the street, "statistics" often means more or less the same as "tables". Statistics Sweden, when it was founded in 1749, was originally called "The Table Agency".

The table concept still dominates very much of the thinking about statistical macrodata, and about metainformation for such data. The strength of this concept is that it is relatively formal, regular, and general. The main weakness of the table concept is that it easily leads to a confusion of the semantical and the syntactical aspects of statistical macrodata, that is, a confusion of the "infological", contents-oriented aspects and the "datalogical", physical and technical aspects.

From a physical point of view a statistical table seems to be basically two-dimensional, since it is usually presented (for example in a publication or on a screen) in terms of rows and columns. For the purposes of this paper, we are not particularly interested in the physical structure of statistical tables.

From a contents-oriented or semantical perspective, a statistical table could be conceptualized as consisting of three or four major dimensions, some of which may be further subdivided into subdimensions. For the time being, we shall associate each one of the major dimensions with a greek letter: α (alfa), ß (beta), γ (gamma), and τ (tau).

2.2 The α ßγ τ - structure of statistical macrodata

The meaning of the four major contents-oriented dimensions of statistical metadata can be briefly described in the following way.

The α dimension

The α dimension of a statistical table is the collective of microlevel objects, which constitutes the **population of objects of interest**, for which some statistical characteristics of interest have been estimated. In business statistics this population is typically some set of enterprises, but it could also be, for example, a set of business or trade transactions of some kind.

The ß dimension

The ß dimension of a statistical table is the set of statistical characteristics, or **parameters of interest**, of the population that have been estimated and presented in the table. The parameters are entities on the macrolevel, but their estimates are computed from microlevel variables of the microlevel objects. The microlevel variables, from which the parameter estimates have been derived, are called ß-variables, and the macrolevel characteristics may also be called ß-characteristics. If the estimated statistical characteristic is a pure frequency, the corresponding microlevel ß-variable will be a "variable" that takes the constant value 1 for every

object in the population.

The procedure used for deriving the estimates of the parameters of interest from the microlevel variables is called an **estimation procedure.**

The γ dimension

The statistical characteristics are usually computed and presented not only for the whole population of interest, but also for subsets of the population, called **domains of interest.** The domains of interest are very often defined in terms of a cross-classification of the population. The microlevel variables that are used for cross-classifying the microlevel objects in the population are called γ-variables, or **classification variables.**

The τ dimension

The τ dimension of a statistical table is the time dimension. The time dimension has an important role in statistical macrodata, not least in economical statistics. However, it is not obvious that time should be regarded as a dimension in its own right. Another alternative is to treat time as an additional γ-variable, but then it should be noted that it is <object, time> pairs that are cross-classified by the extended set of γ-variables, not just individual objects; it should further be noted that summation may not be a meaningful operation along the time axis.

Depending on the ß-variable context, "time" could be interpreted either as *"point of time"* or as *"time period"*. In the context of so-called **stock variables** the former interpretation is the relevant one, whereas the latter interpretation is applicable in the context of so-called **flow variables.**

2.3 The n-dimensional box structure

The cross-classification structure formed by n γ-variables is obviously an n-dimensional structure. It seems justified to regard a statistical table with n γ-variables as an n-dimensional table, a so-called **n-dimensional matrix or box.**

Since the presentation medium for statistical tables is usually two-dimensional, if n > 2, the "logical" dimensions of the table have to be projected onto two "physical" dimensions at presentation time. Most of the γ-variables are usually nested in the so-called stub (the vertical axis) of the presented table, and the rest (if any) have to be nested along the horizontal axis of the presented table.

Thus a box is an n-dimensional structure that is "spanned" by the n γ-variables. It consists of n-dimensional cells, where each cell corresponds to a unique combination of values of the γ-variables. The aggregation process may be thought of as proceeding in the following way:

First each cell of the box contains a number of microlevel objects, all of them having the combination of values of the γ-variables that is determined by the cell's logical place in the box structure, and each of them having a set of values of the ß-variables.

Then the macrolevel estimates of the statistical characteristics will be computed,

cell by cell, by applying an estimation algorithm on the values of the microlevel ß-variables.

Finally, different combinations of macrolevel estimates may be further aggregated over the cells, in order to produce demanded marginal and partial sums. Sometimes the definitions of these "secondary aggregates" are more complicated so that they have to be computed directly from the microdata, rather than from the "primary aggregates"; example: variance estimates.

If the statistical macrodata has a time dimension, this dimension could be included as an (n+1)st dimension of the corresponding box structure.

Hierarchical γ-variables

One or more of the γ-variables in a box structure may be a so-called hierarchical variable, that is, a variable, the values of which are constructed as an m-level code, where a value on level k (k = 1, ..., m-1) could be regarded as an aggregate of a number of values on level k+1; examples: "branch of industry according to ISIC", "type of commodity according to SITC".

If a certain dimension of a box is spanned by a hierarchical variable, partial sums of microlevel ß-variables are formed according to certain typical patterns, when projecting the contents of the box along this dimension.

2.4 "Regular" and "irregular" tables; standardized representation of boxes

In the simplest case there is a one-to-one correspondence between a statistical table, as presented or published, and a box structure according to the definitions above. Such a table will be called a regular table.

In practice it is quite common to edit aggregated statistical data in such a way that one presented table corresponds to a combination of several boxes. Such a table will be called an irregular table.

There may be several reasons for producing irregular tables. For example, an irregular table may be felt to be more compact and/or more readable than an infologically equivalent set of regular tables.

With todays technology it is very simple to produce whatever table presentations that the user would prefer, for example with some desk-top publishing tools, starting from a set of regular tables. Thus for the topic of this paper it is a reasonable simplification to assume that all statistical macrodata can be thought of and described in terms of boxes (or regular statistical tables).

Furthermore, for the future, we could safely assume the existence of some standardized representation format for boxes. Such standardization work is going on at present within the framework of UN/EDIFACT; see reference (4).

One possible standard representation for a box, spanned by n γ-variables (possibly including time), and containing m macrolevel ß-characteristics, could be a so-called flat file or relation (according to the relational data model), where

- the primary key consists of the n γ-variables, and

- the m macrolevel ß-characteristics constitute the remaining, non-key columns of the relational table.

3 Metainformation about statistical macrodata: a proposal for structure and contents

Now we are prepared for a structured discussion of the actual topic of this paper. What metainformation should accompany statistical macrodata, when these data are to be communicated from a producer of statistical macrodata (like a national statistical agency) to some other organization (for example, an international organization like OECD), which will use the data for its own purposes, and/or will make it available, through a statistical database, to other users (like member countries of the organization)? It should be taken into account that the usage of such statistical macrodata typically implies a necessity to compare "the same" data from different producers (that is, different countries).

3.1 Statistical macrodata as a structured list of statistical messages

If we combine some of the ideas of sections 1 and 2 of this paper, we may logically view the macrodata contents of a statistical database as a **structured list of statistical messages**, where a **statistical message** is

- an estimated value of a certain characteristic (or parameter) of interest for a certain collective of objects (population or domain) of interest at a certain point of time (or for a certain time period).

The **logical structuring** of the list of statistical messages can be done in several different ways. One common structuring is to group the statistical messages **by characteristic**. All statistical messages concerning one and the same characteristic will then appear logically close to each other in the list. All the messages in such a group will usually concern **different domains of interest** within (**different time versions** of) **one and the same population** of objects of interest. The different domains of interest (and the different time versions) can usually be defined in terms of one or more box structures, as described in section 2 above. In an international database, "country" will typically be one of the γ-variables spanning the box structures.

If we disregard the actual values in the statistical messages in the structured list, the remaining **skeleton** of the list corresponds to the **tabulation plan** of the statistical database.

On a higher level (than the level of characteristics) the contents of the statistical database (as well as the tabulation plan) may be further structured by criteria like **subject matter topic**, or by (class of) **survey** from which the estimated values of the characteristics emanate.

It should be stressed that the structurings that we are discussing here are purely logical structurings. Thus many different structurings may be defined and used visavi one and the same physical database (with one and the same physical structure). In database terminology, the different structurings correspond to different logical **views**

of the (contents and structure) of the database.

3.2 Metainformation about macrodata and about underlying microdata

Ideally it should not be necessary to describe the underlying microdata in order to describe the meaning and quality of statistical macrodata. In practice this ideal is impossible to attain. In order to achieve a reasonably complete understanding of the meaning and possible usefulness of certain statistical macrodata for different purposes, including the possibilities to compare such data between countries and over time, it is indeed necessary to know something not only about the microdata underlying the macrodata, but also about the procedures that have been used for collecting and processing the microdata, and for computing the estimated macrolevel characteristics.

At Statistics Sweden we are at present developing a new documentation system with the primary goal to facilitate reusage of archived statistical microdata, possibly a long time after the data were originally produced, and possibly by others (like researchers) than those who were themselves engaged in the original production.

3.3 Metainformation that is common for several statistical messages

When we discuss below, what metainformation would be necessary to facilitate proper interpretation and usage of the macrodata in a statistical database, we shall take our starting point in *one* statistical message, that is, one estimated value of one characteristic, for one point (or period) of time, and for one population of objects of interest. Thus we will *think of* the metainformation about macrodata as being organized around individual statistical messages. If this conceptual organization were to be carried out in physical implementation, it would lead to enormous duplication of metadata, since many of the descriptions would be the same for many statistical messages. For example, all statistical messages emanating from the same statistical survey would have a lot of metainformation in common. Even statistical messages emanating from different survey repetitions within one and the same series of survey would have metainformation in common. When such meta-information is to be physically represented in the statistical database, redundant storage should of course be avoided by means of proper physical organization of the metadata. We shall briefly return to this problem in section 5 of this paper, but until then we shall focus our interest on purely contents-oriented problems.

3.4 Proposed metainformation structure

The metainformation associated with a statistical message should contain the information needed by a person, who has no first-hand experience of the survey, underlying the message, but who wishes to interpret and analyze the message and compare it with statistical messages from other sources, for example from other surveys, from other countries, and/or from other repetitions of the same survey.

From our discussion so far we may conclude that this metainformation about a statistical message must include information about

(1) the collective (population or domain) of objects of interest characterized by the message:

   (a) according to intentions and plans, as formulated in a survey design;
   (b) as it actually turned out, in the outcome of the underlying survey;

(2) the characteristic (parameter) of interest, for which the message gives an estimated value:

   (a) according to design;
   (b) according to outcome;

(3) the (point or period of) time, for which the parameter in (2) characterizes the collective of objects in (1):

   (a) according to design;
   (b) according to outcome;

(4) the set(s) of microlevel (observation) objects, for which measurements were made (at certain time points or periods) as a basis for producing the macrolevel message by means of an aggregation process:

   (a) according to design;
   (b) according to outcome;

(5) the set of microlevel (observation) variables, for which values were obtained for the microlevel observation objects in (4):

   (a) according to design;
   (b) according to outcome;

(6) the estimation procedures used when transforming the values of the variables in (5) for the objects in (4) into estimates of the parameters in (2) for the object collectives in (1):

   (a) according to design;
   (b) according to outcome.

We can see three different structures or dimensions in this list of metainformation needs: the micro/macro dimension, the object/property/time dimension, and the design/outcome dimension.

Structure 1: The micro/macro dimension

Microlevel messages, or **observation messages**, reporting observations (measurements) about the values of microlevel variables for microlevel objects at certain times, are used as input to an estimation and aggregation process, producing **macrolevel messages**, or **statistical messages**, as output.

One single macrolevel statistical message, for example the statistical message represented by a number in a single table cell, is typically based on a collection of microlevel observation messages. Sometimes the observation messages will belong

to several different message types, so that on the microlevel there will be several collections of observation messages underlying one single statistical message.

The metainformation needs (4) and (5) above concern the microlevel observation messages, (1), (2), and (3) concern the macrolevel statistical messages, and (6) concerns the estimation and aggregation process.

Structure 2: The object/property/time dimension

Requirements (1) and (4) concern the object (α and γ) parts of messages, requirements (2) and (5) concern the property (ß) parts, and requirement (3) concerns the time (τ) part; cf section 2 of this paper. Requirement (6) concerns the messages as a whole.

Structure 3: The design/outcome dimension

There is both (a) a design and (b) an outcome aspect of each one of the six meta-information requirements.

The three structuring principles are visualized together in figure 1.

4 Structured checklist of the metainformation contents

Figure 1 gives a "skeleton" for the different types of metainformation that could be useful or even necessary to have access to, one way or the other, when one tries to interpret and make use of a certain statistical message, for example a certain figure in a certain cell of a certain statistical table. Now we shall try to add some "meat" to the "skeleton", by listing more specific metainformation kinds that should be considered for different parts of the structure.

By and large the structure of the list will follow figure 1. However there are some minor modifications that will be explained first.

On the microinformation level, we will distinguish between observation messages and derived messages. A **derived message** is a message containing a variable, the value of which has not been directly observed or measured, but rather derived by means of some type of algorithm from values that have been observed or measured. Microlevel derived messages often serve as a kind of **auxiliary messages** between the microlevel observation messages and the ultimately estimated macrolevel statistical messages. One relatively common situation, where the need for auxiliary messages arises, is when the objects of observation belong to another object type than the objects of interest. In such a situation it is not only the variable, but also the object, which is derived.

On the macroinformation level, we shall not particularly emphasize the design/outcome dimension in the list structure. However, the dimension is there, implicitly, in the form of references to microlevel metainformation, where the distinction is more visible.

Finally, in the listing below, the metainformation concerning the estimation procedure will be logically placed together with the property-oriented meta-information about the statistical message.

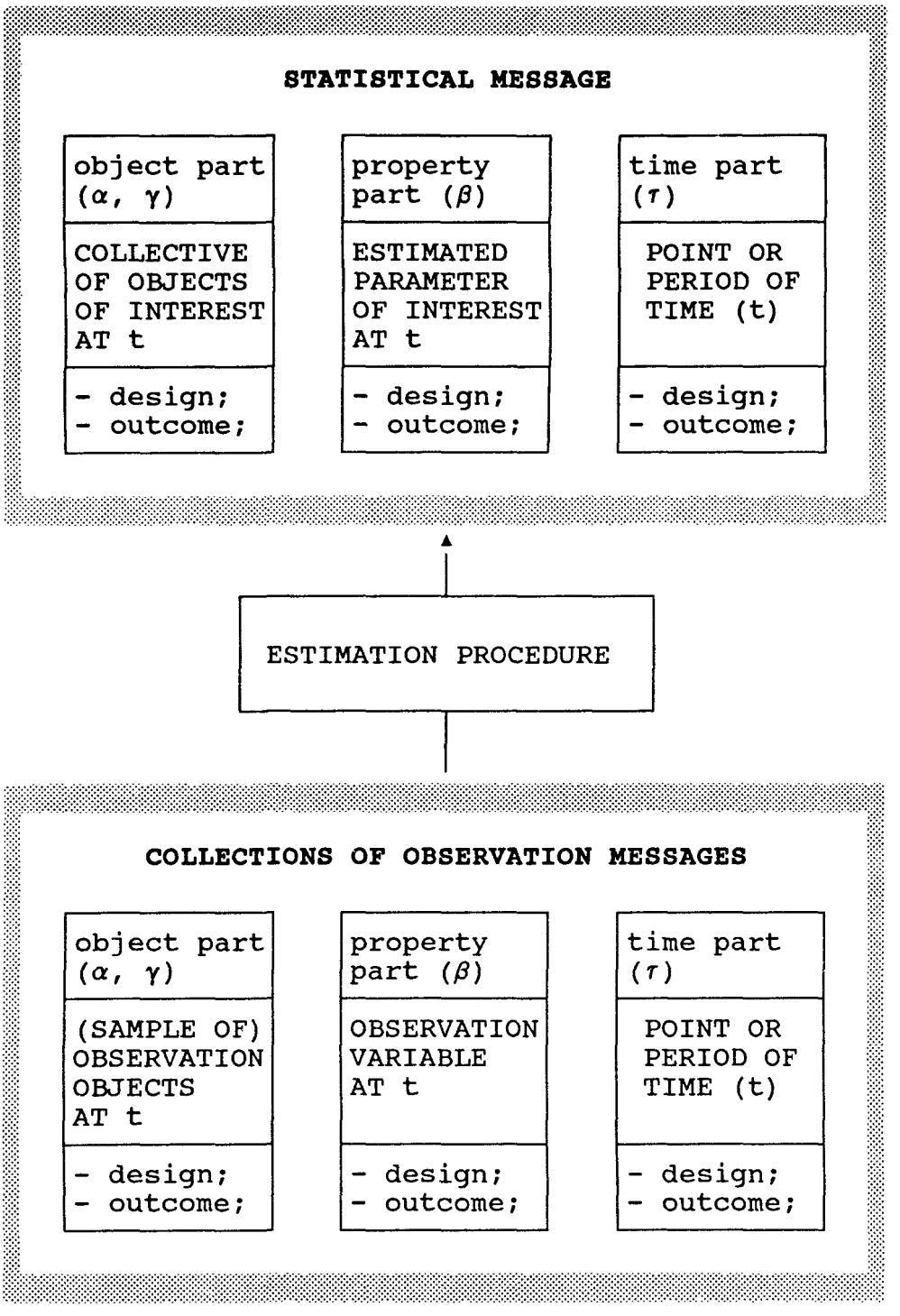

Figure 1. *Proposed metainformation structure for statistical messages.*

4.1 Metainformation about microlevel observation messages

- Identification of microlevel message type;

  - name of object type;
  - name of variable;
  - time scale and time value;

4.1.1 Object-oriented metainformation about observation messages

- Objects of observation: name of object type and short description;

Intended objects of observation according to design

- Definition of object type;

  - existence (birth and death) criteria;
  - identification criteria;

- Definition of population and frame;

  - definition of population;
  - definition of frame and frame procedure (linkage between frame elements, respondents, and population objects);
  - overcoverage and undercoverage;

- Definition of sampling procedure (if applicable);

- How does the design of the object-oriented part of this message type relate to international standards, agreements, and recommendations?

Actual observation objects according to outcome

- Deviations from the design as regards definitions of object type, population, frame, frame procedure, and sampling procedure;

- Non-response objects: description, actions, treatment;

4.1.2 Property-oriented metainformation about observation messages

- Observation variable: name and short description;

Intended variable of observation according to design

- Definition of the intended meaning of the observation variable;

- Design of the measurement of the observation variable;

  - measurement method and measurement instrument;
  - question asked (or equivalent);
  - possible response values, measurement scale;

- Rules for coding and editing;
- How does the property-oriented part of this message type relate to international standards, agreements, and recommendations?

Actual variable observations according to outcome

- Experiences of the measurement method and measurement instrument;
- Experiences of coding and editing;
- Non-response: description, actions, treatment;

4.1.3 Time-oriented metainformation about observation messages

About the intended measurements, according to design

- Object system time: time scale and time value(s) of the phenomena, which are reported by the messages belonging to the message type;
- Information system time: time period, during which the measurements were planned to be carried out;
- Changes in the design of the underlying survey, which occurred at this time, and which affected the definitions etc stated in the object- and/or property oriented parts of the metainformation for this message type; causes of such design changes;
- How does the time-oriented part of this message type relate to international standards, agreements, and recommendations?

About the actual measurements, according to outcome

- Actual object system time (if different from planned time);
- Actual measurement time (if different from planned time);
- Exceptional events (labour market conflicts, earthquakes, etc) that occurred during the measurement time period, and that may have affected the measurements;
- Information concerning seasonal (or similar) variation that may be of relevance for seasonal (or similar) adjustments;

4.2 Metainformation about microlevel derived messages

As was mentioned earlier, we sometimes need intermediate messages, which are derivable from observed messages, in producing a certain statistical message of interest. Both the object part and the property part of such a message may be derived, but sometimes it is only the property part (or the object part), which is derived; the other part is then observed.

In principle the metainformation about derived messages should be more or less the same as the metainformation about observed messages, with the exception that everything that has to do with the measurement will be replaced by formal definitions in terms of other messages, observed or derived. We will not go into the details here.

4.3 Metainformation about macrolevel statistical messages

- Identification of macrolevel message;
  - name of collective of objects;
  - name of parameter;
  - time scale and time value;

4.3.1 Object-oriented metainformation about statistical messages

- Name of collective of objects (population or other domain of interest);
- Definition of collective of interest in terms of microlevel message types, including references to (cf section 2)
  - microlevel object type;
  - microlevel population of objects;
  - microlevel α-selection property, if applicable;
  - microlevel γ-variable-based selection property, if applicable;
- How does the object-oriented part of this statistical message relate to international standards, agreements, and recommendations?

4.3.2 Property-oriented metainformation about statistical messages

- Name of estimated parameter;
- Specification of estimated parameter, including
  - verbal description;
  - range of values, measurement scale;
- Definition of estimated parameter in terms of microlevel message types, including references to (cf section 2)
  - microlevel ß-variables;
  - aggregation function;
  - other transformation functions, e g for seasonal adjustment, transformation into normalized scales like fixed-prices, etc;
- Estimation procedure for the estimated parameter;
  - definition of the estimator, taking into account the sampling design for the underlying survey, non-response, and other uncertainties and errors that are compensated for; weights, weighing procedures, imputation rules;

- Estimation of the uncertainty of the parameter estimate;

  - definition of the uncertainty estimator;
  - evaluated value of the uncertainty estimator for this statistical message; (for example a so-called "variance-estimate");

- How does the property-oriented part of this statistical message relate to international standards, agreements, and recommendations?

4.3.3 Time-oriented metainformation about statistical messages

- Object system time;

- Measurement time;

- References to time-related transformation functions applied in producing this statistical message; for example:

  - method for seasonal (and similar) adjustment;
  - base year (for index)

- How does the time-oriented part of this message relate to international standards, agreements, and recommendations? For example, is a statistical message of this type produced for prescribed time periods, at prescribed time intervals, and with prescribed maximum reporting delay?

5 Ambition levels and simplification possibilities

The metainformation checklist generated by the discussion in this paper may seem to be very ambitious, especially if we imagine that every individual statistical message in a database of statistical macrodata should be accompanied by all the metainformation items identified above. In fact, the checklist *is* ambitious. It should be interpreted as a *maximal list*. If all the metainformation indicated in the checklist were available to a user of a statistical database, there is a good chance that the user, provided that he or she has the general education and training that one could expect, could actually interpret the data more or less correctly, and analyze them and compare them with similar data from other sources in a responsible way. This may not be quite easy to *prove*, but at least we can try to convince ourselves by going back to the four requirements stated in section 1.2 of this paper, and checking that these requirements are matched by metainformation items in the checklist. This is left to the reader as an exercise.

There are several obvious possibilities to reduce potential redundancy when storing the metainformation for a whole database of statistical macrodata. In section 2 of this paper we discussed tables and boxes of macrodata. Every **cell** in a table or box has a one-to-one relationship to a statistical message. Thus a single table or box may easily contain hundreds or thousands of statistical messages, and to give all the metainformation in our checklist message by message (cell by cell) would require much more space than the table or box itself. On the other hand we could obviously give many of the metainformation items in our checklist for many messages (cells) at the same time. In many cases the only metainformation differences between the different messages in a table or box will concern the domains of interest; every cell

corresponds to a uniquely determined domain of interest within a population, which is the same for all the cells in the box. Furthermore, even if the domain of interest is unique for every cell, all the domains of interest in the box can usually be defined in terms of the same combination of (maybe two, three, or four) γ-variables.

As an example, consider a box labeled *"Value of export from Sweden 1990 by commodity types (according to SITC) and country of destination"*. The number of cells (and hence the number of statistical messages) in this box will be equal to

(the number of commodity types) × (the number of countries of destination);

However, all these messages will be aggregated from the same set of measurements in one and the same survey, the Swedish foreign trade statistics, 1990. There will be one object type (and one population of objects) underlying all the observations (Swedish export transactions) and one observation variable (value). Only the domain of interest (within the population) will vary from message to message, from cell to cell. But in order to give a generic description of all the domains of interest it is sufficient to describe the two microlevel observation message types, corresponding to the two γ-variables, "commodity type" and "country of destination".

As is illustrated by this example, the metainformation for thousands of statistical messages can sometimes be given in a very compact way, provided that the messages belong to one and the same box, emanating from one and the same survey. There may even be metainformation, which is common for statistical messages in different boxes and/or from different surveys. For example, there may be common definitions and descriptions of

- object types;
- populations;
- variables;
- estimators;
- methods for seasonal adjustment.

Entities such as these, for which there is metainformation, which is common for many individual statistical messages, should be regarded as **metainformation objects**, or **metaobjects**, in a **conceptual model** for the metainformation subsystem, the so-called **metadatabase**, of a statistical database.

On the other hand, it should also be noted that even if there are often obvious and attractive possibilities to "factor out" common metainformation items for many statistical messages, the simplification possibilities will not be the same for all surveys and all message collections. Experience shows that one will often need a lot of flexibility in metainformation systems, on the input side, as well as on the output side. Thus sometimes it will even be necessary to make use of the full complexity of the checklist presented in this paper, by giving, for a single individual statistical message in the database, a unique metadata description for every individual item in the checklist.

6 References

1. Rauch, Lars: *"Metadata in the Statistical Databases of Statistics Sweden"*, Statistics Sweden 1991.

2. United Nations / Economic Commission for Europe: *"AXIS Reference Manual"*, Geneva 1990.

3. Nordbäck, Lars: *"An Illustrated Booklet on the Dissemination of Statistics in the 1990s"*, Statistics Sweden 1990.

4. United Nations / Economic Commission for Europe: *"MD6 Statistics of the Western European EDIFACT Board"*, Geneva 1990; prepared by Philippe Lebaube, EUROSTAT, chairman of MD6.

5. Malmborg, Erik: *"SCB-MATRIX, A Proposal for Data and Metadata Syntax for Statistical Data"*, Statistics Sweden 1990.

6. Sundgren, Bo: *"Statistical Databases and Statistical Information Systems"*, Lecture notes, Statistics Sweden 1990.

7. Sundgren, Bo: *"Some Properties of Statistical Information: Pragmatics, Semantics, and Syntactics"*, Statistics Sweden 1991.

R & D Reports är en för U/ADB och U/STM gemensam publikationsserie, som fr o m 1988-01-01 ersätter de tidigare "gula" och "gröna" serierna. I serien ingår även **Abstracts** (sammanfattning av metodrapporter från SCB).

R & D Reports Statistics Sweden are published by the Department of Research & Development within Statistics Sweden. Reports dealing with statistical methods have green (grön) covers. Reports dealing with EDP methods have yellow (gul) covers. In addition, abstracts are published three times a year (light brown /beige/ covers).

Reports published during 1991:

1991:1 (grön)	Computing elementary aggregates in the Swedish consumer price index **(Jörgen Dalén)**
1991:2 (grön)	Översikt av estimatorer för stora och små redovisningsgrupper **(Sixten Lundström)**
1991:3 (grön)	Interacting nonresponse and response errors **(Håkan L Lindström)**
1991:4 (grön)	Effects of nonresponse on survey estimates in the analysis of competing exponential risks **(Ingrid Lyberg)**
1991:5 (grön)	Study of the estimation precision in the Chinese MIS surveys **(Bengt Rosén)**
1991:6 (beige)	Abstracts I - Sammanfattning av metodrapporter från SCB
1991:7 (grön)	Kvalitetsfonden 1990: Projekt - handläggning - utvärdering **(Roland Friberg och Håkan L Lindström)**
1991:8 (grön)	Tre återintervjustudier: Inkomstfördelningsundersökningen **(Jan Eriksson)**, Högskoleenkätuppföljningen **(Anders Karlsson)**, Årsarbetskraften **(Majvor Karlsson och Solveig Thudin)**

Kvarvarande **beige** och **gröna** exemplar av ovanstående promemorior kan rekvireras från Inga-Lill Pettersson, U/LEDN, SCB, 115 81 STOCKHOLM, eller per telefon 08-783 49 56.

Kvarvarande **gula** exemplar kan rekvireras från Ingvar Andersson, U/ADB, SCB, 115 81 STOCKHOLM, eller per telefon 08-783 41 47.