Statistiska centralbyrån Statistics Sweden

Our Legacy to Future Generations Using Databases for Better Availability and Documentation

Gösta Guteland and Erik Malmborg

R&D Report
Research - Methods - Development
1996:6

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 1996:6. Our legacy to future generations using databases for better availability and documentation / Gösta Guteland; Erik Malmborg.
Digitaliserad av Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-1996-X101OP9606

Statistiska centralbyrån Statistics Sweden

Our Legacy to Future Generations Using Databases for Better Availability and Documentation

Gösta Guteland and Erik Malmborg

R&D Report
Research - Methods - Development
1996:6

Från trycket Januari 1997
Producent Statistiska centralbyrån, utvecklingsavdelningen
Ansvarig utgivare Lars Lyberg

Förfrågningar Erik Malmborg
telefon 08-783 40 27
telefax 08-783 45 99

ISSN 0283-8680

Printed January, 1997
Producer Statistics Sweden, Department of Research and Development
Publisher Lars Lyberg

Inquiries Erik Malmborg
telephone +46 8 783 40 27
telefax +46 8 783 45 99

Paper presented at Seminar on Official Statistics - Past and Future, Lisbon, Portugal, September 1996

ISSN 0283-8680

Our legacy to future generations - using databases for better availability and documentation

Gösta Guteland and Erik Malmborg
Statistics Sweden

1. Introduction

We have a long tradition in Sweden of producing statistics - our regular demographic statistics dating back as far as 1749. These older statistics were very carefully documented, which has facilitated a vast amount of research. It is still possible to read the original forms, and browsing the material is easy. The data can be used to analyse causes of death, educational levels in the population, mortality, and a range of other things.

The question now is: Are *we* going to leave a documentation of our society that is of as high a quality as they did 250 years ago? Will it be possible to read all our electronic documents in another 250 years? Is our documentation good enough? Are our methods of storage safe enough?

Great technological strides have been made in our production processes since 1749. Data collection has been immensely facilitated by the introduction of postal and telephone services, as data preparation later on has been made easier by computer technology. Statistics Sweden invested in its first mainframe at the beginning of the 1960s. The real problems with documentation date from that time on. We have tapes from the 1960s that we can't read anymore. Fortunately enough we had saved the forms, so that it is still possible to reconstruct the statistical tables.

2. From mainframes to PC networks

Statistics Sweden, like many other NSIs, is now replacing its mainframe with local networks. This has been a time-consuming process. We started the development of a computerised office information system several years ago. Recently, we have begun converting to a new, PC-based platform for our production, a task we expect to be completed at the end of 1999. By that time we will be carrying out most of our production in the new environment. Why the length of time? One reason is that data have not always been stored in a proper manner and there may be shortcomings in the documentation. Another reason is the lack of people able to work with the new techniques.

Given the difficulties, what then are the advantages of the decision to change over to a new technical platform? Above all, the possibilities for finding cheaper and more flexible solutions. With the new platform we will be able to make very rapid manipulations in PCs

and servers, and data can be transported to users in a form that allows them to make their own calculations.

Our new system for production and distribution is illustrated in the following figure.

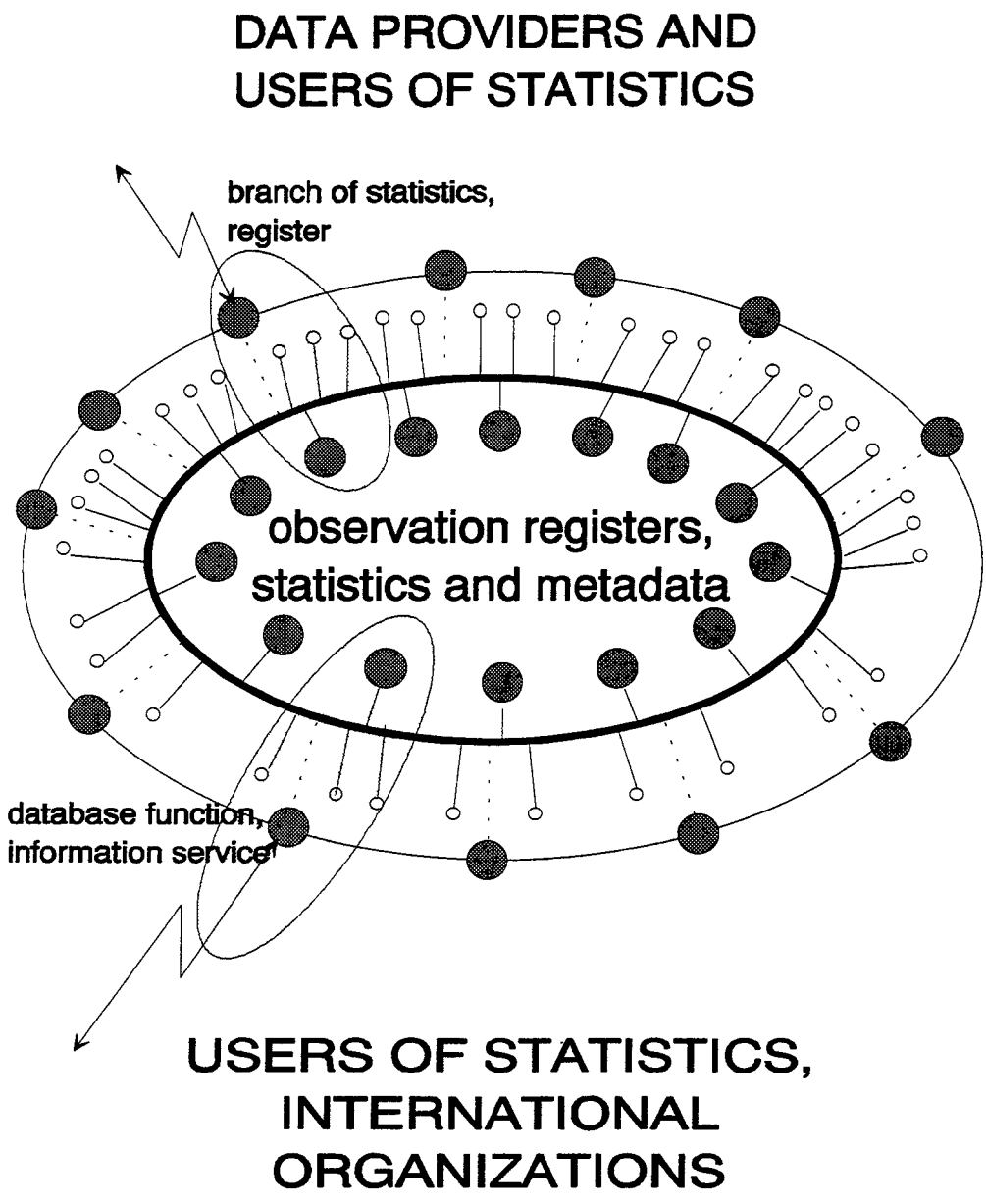

Fig. 1

The inner circle consists of PCs (small dots) and servers (big dots) loaded with statistics and raw data that only those working at Statistics Sweden are authorized to use. Most data are stored on the servers. Metadata or data on data are also included and the servers are linked to each other in such a way that it is possible to combine data from different servers.

The outer circle contains macrodata, metadata and some microdata divested of any information that would make it possible to identify a certain person or enterprise. Users of statistics will be able to reach the outer circle via the Internet or some other on-line connection, but there will be no possibility of straying from the outer to the inner circle, as all links are equipped with firewalls or other safe constructions.

Of course, not all users will have to connect on line in order to get at our statistics. A floppy disk or CD-Rom can be sent instead. As before, much of our results will also be published in books and reports, and we will continue to send our press releases to the media.

Given the complicated nature of the system it is easy to understand that good quality documentation may be difficult to maintain, compared to the rather simple forms and tables that were produced 250 years ago, or even compared with the mainframe production system of 35 years ago.

3. Databases with official statistics

Everything cannot be saved for the future. Even if we had very large computers we could still not store everything. A very important question therefore is: What shall we save for future generations? Should we store all primary data (microdata)? And to what extent should we store aggregated data (statistics). Shall we save all the statistics designated as "official" or only a part of it? How are "official statistics" to be delimited? And what other types of statistics should be saved?

Sweden has special laws and regulations concerning what must be included in our national archives. These details are not appropriate here. For us what is most important at this time is to work out how our material should be handled in order to facilitate long-term filing.

"Official statistics" in Sweden were not given a particular definition until 1994, in connection with a statistical reform giving a number of other government agencies responsibility for some of the statistics formerly within Statistics Sweden's sphere. Until then, "official statistics" was simply the designation of a publication series. In order to separate the statistics that were to remain with Statistics Sweden from other statistics it was necessary to arrive at an exact definition of what is meant by "official statistics". Such a list was therefore drawn up and adopted by the Government.

For the new official statistics special rules are given. Documentation must be according to instructions prepared by Statistics Sweden. All agencies with responsibility for official statistics are obliged to submit detailed descriptions of their products to Statistics Sweden, which are then placed in a metadatabase, accessible to anyone seeking information about official statistics.

From the metadatabase it will be possible to proceed to the actual statistical tables. By clicking in the metadatabase on, say, Labour Market Statistics, one can then come to Labour Force Statistics or statistics on any other part of the labour market. In the metadatabase information is also given on the quality of the statistics as well as such relevant information as coverage and so on.

The new technical platform will form the foundation for our new databases of official statistics. However all official statistics will not be placed in them, but only such as are of wider interest, primarily those statistics in general demand by users in different sectors of the society. These statistics, at least, will definitely be well documented and saved for future generations. A number of registers (e.g. population, enterprises) and some "observation registers" from important surveys will also be part of the databases. An observation register is the final edited register with the observations from the survey. An important aspect discussed at some length later in the paper is the documentation needed for these registers.

4. Documentation and metadata

As described above, documentation or "metadata" is an important part of our new databases. A main theme of this paper is the need for systematic handling of metadata. Metadata is a technical term for data describing other data. The data thus described are the major products of statistical work, i.e. statistics (tables, time-series, etc.), but also "micro-data" or "observation registers". As can be seen, a specific terminology has developed, of which some important parts will be presented here. The reader wanting a more complete presentation is referred to the "Guidelines for the Modelling of Statistical Data and Metadata" (see References). These guidelines, edited by Prof. Bo Sundgren of Statistics Sweden, are a result of the UN/ECE METIS project. Prof. Sundgren is also the project leader for the development of the new Swedish databases. It is therefore natural for our database development to follow these guidelines.

A development project with strong influence on the guidelines was started at Statistics Sweden in 1990. Prof. Bengt Rosén (statistics) and prof. Bo Sundgren (statistical informatics) together developed a new documentation model for statistical surveys. The following are some important conclusions from their work:

- Documentations for production systems, observation registers and statistical databases have a lot in common. There are, for example, strong arguments for documenting such things as sampling, data editing and estimation when archiving an observation register.

- Documentations for observation registers and for statistical databases can partly be derived from documentations for the survey production systems.

- When documenting a user-oriented statistical database, there is a need for contributions from several different production systems.

- Metadata should only be registered once, and then reused. The natural situations for this registration are when developing the production systems, and when using them for production of statistics.

5. The needs of the future

Researchers often want to reuse data from surveys made at an earlier point in time. Often the original survey was made several decades earlier, as might be the case when searching for the causes of e.g. some form of cancer. It is indeed a difficult problem to reuse survey data:

- It might be impossible to access the staff who once (maybe decades earlier) produced the data.

- The data typically is used for another purpose than the purpose of the original survey. It is then important to be able to assess the assumptions originally made when designing the survey. Important aspects are the models used for sampling, data editing and estimation.

We can see the importance of keeping detailed information on the processes needed to create the observation register. The documentation template for observation registers as given in the guidelines is reproduced in Fig 2.

The documentation model is designed to handle repetitive as well as one-time surveys. A repetitive survey, or a survey series, consists of several similar surveys, e.g. producing new values for some economic indicator (e.g. price indices). The design of the individual surveys in a survey series is basically the same, but might be modified over time. Moreover, each data collection may have its own problems with e.g. non-response rates.

The complete documentation model can be found in the UN/ECE Guidelines. It consists of background material and templates for:

- Quality Declaration of Statistical Data
- Observation Register Documentation
- Documentation Template for a Statistical Survey

The templates are interrelated but have different purposes. Typically the documentation for a survey series is updated when modifications are made to the survey design. These changes typically do not influence quality declarations or observation register documentations from earlier surveys in the series, but do influence the documentations from later surveys in the series.

We have chosen to show the observation register documentation template in this paper, as it is very important for the long-term reuse of data from statistical surveys.

OBSERVATION REGISTER DOCUMENTATION

0 Administrative information

0.0 Documentation templet
0.1 Survey name and identification, organisation and persons responsible
0.2 Documentation modules and subsystems
0.3 Archived data sets and published statistics
0.4 References to other relevant documentation

1 Survey contents

1.1 Domain of interest and target domain, verbal description
1.2 Target domain, formal description
- 1.2.1 Target objects, description and object graph
- 1.2.2 Target populations
- 1.2.3 Target variables

1.3 Survey outputs
- 1.3.1 Structured overview of the tabulation plan
- 1.3.2 Publications in printed form
- 1.3.3 Electronical distribution
- 1.3.4 Database storage

2 Survey plan

2.1 Frame procedure and observation objects
- 2.1.1 Overview
- 2.1.2 Frame and its links to objects
- 2.1.3 Frame production
- 2.1.4 Overcoverage and undercoverage

2.2 Sampling procedure (if applicable)
2.3 Data collection procedure
- 2.3.1 Observation objects, description and object graph
- 2.3.2 Data sources, including contact procedures
- 2.3.3 Observation variables and measurement instruments
- 2.3.4 Interruptions (including actions at overcoverage)
- 2.3.5 Non-response actions

2.4 Planned data preparation (coding, data entry, editing and correction)
2.5 Planned observation register
- 2.5.1 Overview
- 2.5.2 Object types, including derived object types
- 2.5.3 Object graph
- 2.5.4 Object/variable-matrixes, including derived variables
- 2.5.5 Data set descriptions
- 2.5.6 Derivation procedures (in complicated cases)

3 Completed data collection

3.1 Frame production
3.2 Sampling
3.3 Data collection
- 3.3.1 Communication with the data providers
- 3.3.2 Measurements, experiences of instruments
- 3.3.3 Interruptions/overcoverage, actions taken
- 3.3.4 Non-response, causes and actions taken
- 3.3.5 Editing and correction at data collection time

3.4 Data preparation (coding, data entry, editing and correction)
3.5 Production of final observation register
- 3.5.1 Treatment of interruption/overcoverage objects
- 3.5.2 Treatment of non-response objects
- 3.5.3 Treatment of partial non-response
- 3.5.4 Frequency counts of overcoverage, responses, non-responses etc
- 3.5.5 Completed derivations of derived objects and variables

4 Statistical processing and presentation

4.1 Observation models
- 4.1.1 Sampling
- 4.1.2 Non-response
- 4.1.3 Measurement/observation
- 4.1.4 Frame coverage
- 4.1.5 Total model

4.2 Population models
4.3 Computation formulae for estimations
- 4.3.1 Point estimations
- 4.3.2 Estimations of sampling errors (variance estimations)
- 4.3.3 Estimation/judgment of other quality characteristics

4.4 Analyses
4.5 Presentation and dissemination procedures

5 -

6 Log-book

Fig. 2

6. Practical consequences for the organisation of statistical work

Many statistical metadata systems in the past have failed for various reasons:

- Metadata collection is dull, expensive, and time-consuming
- The natural providers of metadata are distinct from the typical users of metadata. The providers typically do not feel motivated to take on the burden of creating formalized metadata. They already have the knowledge. The users, on the other hand, need but cannot themselves produce the metadata.
- Users of statistical data are usually interested in several (related) collections of data, which have been collected by different surveys. They will thus not find a metadata system to be of much value, until the metadatabase covers data from many surveys.

Huge metadata collection activities should be avoided. Instead, as much as possible of the metadata should be generated as a side effect of other activities.

Using computer support (c.f. Malmborg/Lisagor) for documentation, it is feasible to have an evolving documentation for the production system of the survey series. For each survey ("survey round") a "snapshot" of this documentation is the base for the observation register documentation. Special information for the individual survey (e.g. non-response rates) must be added. In the documentation template of Fig 2, Section 3 typically must be revised for each survey round.

7. Metadata handling in the new Swedish databases

Databases with macro-data (matrix and time-series data) have a long tradition within Statistics Sweden. Our main-frame database system AXIS has been used within several organisations internationally (including UN/ECE). Our new databases are to continue this tradition, but also add several new benefits:

- More descriptive metadata of quality type will be available. The "open architecture" of the new systems will make metadata more available for different types of use.
- Observation registers from some important surveys and some "base registers" (e.g. the population register, the enterprise register) will be made more easily available for researchers and investigators. Of course, full control will be retained concerning privacy for individual objects.

Earlier statistical databases typically have some limited amount of descriptive metadata, e.g. in the form of footnotes. The new databases in Sweden will have a text-based metadata base containing a lot of information:

- Product descriptions for all offical statistics in Sweden. This includes quality information from the different surveys.
- Observation register documentations for important registers.
- Publication plans for official statistics.
- Information on archived registers.
- ...

These text-based databases can be used for finding information on available statistics. An important aspect is the couplings between this text-based metadatabase and the more formalized databases with statistical data. Special links are developed to handle some important situations:

- If the user is retrieving statistical data from a survey, s/he might be interested in additional descriptive metadata on the survey. This is reached by "pointers" to relevant documents with quality information.

- If a user has found out that there is survey data of interest in a data base, s/he will want a quick link to the published data to be able to retrieve it.

8. Consequences for the European cooperation

Eurostat, the statistical office of the EU, has initiated an ambitious project for data exchange between member states. These "Distributed Statistical Information Services" (DSIS) naturally depend a lot on metadata. To handle these aspects, a "Meta Data Task Force" has been set up as part of the project. This Task Force is to:

- Propose what should be included as metadata for the purposes of DSIS.
- Establish a data model and propose a metadata exchange format. The format proposed is the GESMES/ECOSER format from the UN/EDIFACT work (see References).
- Contribute to the requirements specification of the Master Metadata Service (which is a central part of the DSIS architecture).

The metadata handled by the Master Metadata Service are typically of the type documenting aggregated data and time series. The level of ambition must be kept on a reasonable level, as the metadata shall be translated into several languages and this is an expensive process.

As was argued above, comparative use of statistical data in new situations sometimes require metadata of a more process oriented type. It would be unreasonably ambitious to include all this metadata (c.f. the documentation templates in the guidelines). On the other hand there might be situations where this metadata is very important. A pragmatic approach to handling this situation, proposed by Sweden and Denmark in the Metadata Task Force, is to include formal links in the Master Metadata Service to the national documentation systems. These systems can have different levels of ambition and different structures, but with the links further information can be found if really needed.

9. Our legacy to the future

We have tried to show how a systematic handling of metadata can help to provide present and future users of data with the needed background information. The research type user trying to resuse data from past surveys has special needs that must be catered for. A key issue is having links between different types of metadata.

A future user of statistical data (including archived observation registers) must be able to find information about the present survey -processing routines.

The building of the new Swedish statistical databases is made with these concerns in mind. It is our hope that the future researcher using results from present-day surveys shall be in a still better position than the researchers using Swedish 18th century demographic statistics.

References

UN/ECE (1995) *Guidelines for the Modelling of Statistical Data and Metadata.* Report from the UN/ECE METIS Group, established within the programme of work of the Conference of European Statisticians. Basically the same text is available from Statistics Sweden as:

Sundgren, B. (1994) *Statistical Metadata and Metainformation Systems - an update.* Report for the UN/ECE METIS Group.

Malmborg, E. and Lisagor, L. (1993) *Implementing a Statistical Meta-Information System* Proceedings of the Statistical Metainformation Systems Workshop in Luxemburg, February 1993. Also in Statistical Journal of the United Nations UN/ECE 2/1993. Also available from Statistics Sweden.

Malmborg, E. and Sundgren, B. (1994) *Integration of statistical information systems - theory and practice* Proceedings of the Seventh International Conference on Scientific and Statistical Database Management, University of Virginia, USA, September 1994, IEEE Computer Society Press

UN Western European EDIFACT Board, Message Development Group 6 Statistics (1995) *GESMES/ECOSER User Guide.* Published by Eurostat, Luxemburg

Förteckning över utkomna R&D Reports

R&D Reports är en för U/ADB och U/STM gemensam publikationsserie, som 1988-01-01 ersatte de tidigare "gula" och "gröna" serierna. I serien ingick fram till årsskiftet 1992-93 även **Abstracts** (sammanfattning av metodrapporter från SCB).

Reports published during 1996:

1996:1 (grön) On Sampling with Probability Proportional to Size *(Bengt Rosén)*

1996:2 (grön) Bortfallsbarometern nr 11 *(Antti Ahtiainen, Stefan Berg, Margareta Eriksson, Åsa Greijer, Dan Hedlin, Monica Rennermalm, Anita Ullberg)*

1996:3 (grön) Regression Estimators in Theory and in Practice *(Tomas Garås)*

1996:4 (gul) Quality Aspects of a Modern Database Service *(Pat Dean and Bo Sundgren)*

1996:5 (gul) Metadata: A Quality Element in Official Statistics - the Swedish Approach *(Bo Sundgren and Pat Dean)*

Tidigare utgivna *R & D Reports* kan beställas genom Ingvar Andersson, SCB, U/SIB, Box 24300, 115 81 STOCKHOLM (telefon 08-783 41 47, fax 08-783 45 99, e-post ingvar andersson@scb.se).

R & D Reports listed above as well as issues from 1988-1994 can - in case they are still in stock - be ordered from Statistics Sweden, att. Ingvar Andersson U/SIB, Box 24300, S-115 81 STOCKHOLM (telephone +46 8 783 41 47, fax +46 8 783 45 99, e-mail ingvar andersson@scb.se).