STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

COMMISSION OF THE EUROPEAN COMMUNITIES

CONFERENCE OF EUROPEAN STATISTICIANS

EUROSTAT

Joint UNECE/Eurostat Work Session on Statistical Metadata
(6 - 8 March 2002, Luxembourg)

Working Paper No. 9
English only

Topic (i): Infrastructure issues for statistical metadata

THE STRUCTURE OF THE METADATA SYSTEM AT STATISTICS SWEDEN

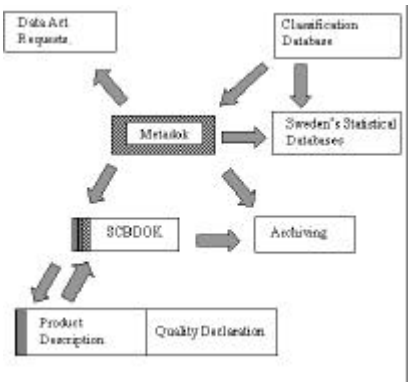
Submitted by Statistics Sweden[1]

Contributed paper

ABSTRACT

The importance of metadata for internal as well as external users has been stressed at Statistics Sweden (SCB) for many years. Despite this fact, the actual documentation work has been rather slow and in 2001, a comprehensive project was initialised in order to speed up documentation work. To obtain a basis for future development, an extensive mapping of the needs and wishes among those concerned has been carried out. In the short run, training and user's guides as well as more stable tools have been given priority.

The diagram on the right illustrates the meta data system of today with the arrows showing how the different parts are linked together. A basic thought is that once information has entered the system, it can be used again when the same information is asked for in other parts of the system. The broken line areas show how information in one part of the system can be copied to and directly reused in other parts.



There are three kinds of documentation tools and templates that are part of the metadata system at SCB:

?? *Product description*: a template for short information about quality and other basic facts for all official statistics.

?? *SCBDOK*: a template for detailed descriptions of how a survey is conducted from data collection to dissemination for all observation registers and productions systems, for which SCB is responsible.

?? *Metadok*: a software tool used to describe the contents of an observation register, which stores formalised metadata that can be used by other software tools.

There are a number of possibilities to benefit from completed documentations. All three kinds of documentation are presented on the Internet and Intranet, enabling external and internal users to search among statistical products and registers. Automatic routines use Metadok metadata for archiving and for

[1] Prepared by Elisabet Andersson and Linda Ryen.

the production of so-called data act requests (extracts showing what data is stored at SCB about a person, if this person has submitted an application for this).

Our conclusion is that three factors are very important for documentation and the metadata system: technology, knowledge and attitude.

?? *Technology*: In order to make the system more functional and user friendly, further deployment of the already existing parts and connections, as well as of new applications, is necessary. Better possibilities to use metadata during the production process are considered important.

?? *Knowledge*: There is a great need for education in the metadata area about the available tools and templates and about how the parts in the metadata system are connected. In autumn 2001, employees were offered one-day documentation courses in order to get started with their documentation. In total, around 150 persons were trained on 13 occasions.

?? *Attitude*: Working with documentation and metadata is traditionally considered as boring and has a low status. Changing attitudes is a time-consuming task but it is necessary in order to make it possible to attain the goals that are set within the metadata area.

I. INTRODUCTION

1. The importance of metadata has been stressed at Statistics Sweden (SCB) for many years and the concept "metadata" is familiar to most of the employees. Despite this fact, the actual documentation work has been rather slow. There are several reasons for this, but the fundamental reason is that the metadata field has suffered from inadequate resources. The existing software tools have not been sufficient to fulfil the needs of the subject area departments. Because of this, it has been difficult to obtain the necessary metadata from the departments.

2. Metadata is important for external as well as internal users. The former need metadata to make correct interpretations and analyses of the statistics from SCB. Metadata richness is also crucial for researchers both today and in the future. From an internal point of view, metadata is of great importance as a support for the producers of statistics. A well-functioning metadata system is a precondition for more efficient production, which in turn becomes less vulnerable to staff turnover.

3. The subject-area departments are responsible for the production of metadata at SCB. According to requirements set by the Director General, the departments must complete three kinds of documentation in 2001: Product descriptions, Metadok and SCBDOK, described in more detail below. First and foremost, this concerns 17 important observation registers, which have been especially appointed by the Swedish government. The documentation of the remaining Official Statistics is also considered to be of very high priority.

II. THE METADATA PROJECT

4. The metadata field is now given high priority at SCB, and management declared 2001 as "the year of documentation". A comprehensive project has been initialised in order to speed up documentation work while emphasizing the importance of metadata. In the long term, SCB strives to attain an efficient and user-friendly metadata system. To obtain a basis for future development, an extensive mapping of the needs and wishes among those concerned has been carried out. In the short run, training and user's guides as well as more stable tools have been given priority.

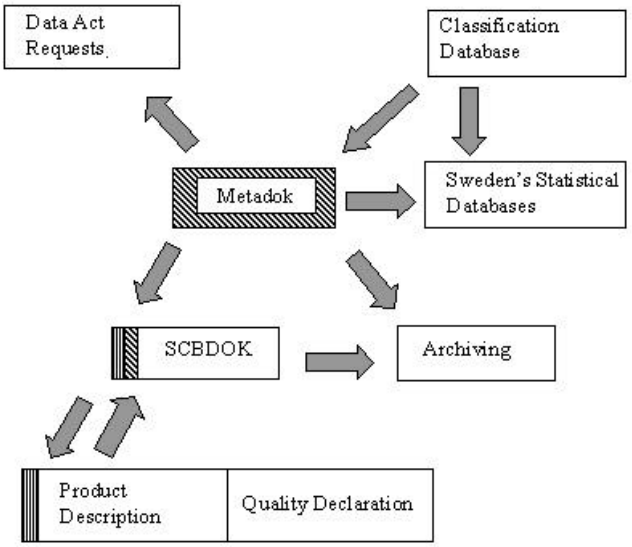
5. In the first six months much effort was made to improve the available tools. It is crucial that existing programmes function well, because otherwise it is unreasonable to require documentation from personnel. In autumn 2001, employees were offered one-day documentation courses in order to get

started with their documentation. In total, around 150 persons were trained on 13 occasions. The courses included demonstrations of the tools/templates and their applications, and exercises were carried out where the participants began to document their own registers with support from the teachers. According to those responsible in the subject-area departments, one of the main difficulties with documentation was getting started. For this reason, we considered it important that the participants received support when documenting their own registers, and were able to ask questions about contents and delimitations. A meeting completed the course where deficiencies in the present metadata system and development possibilities were discussed. The viewpoints presented during the courses are a very valuable input for future development of the metadata system. A general user's guide to the metadata system at SCB has been written especially for the courses, explaining how the system is built-up.

6. The project has also tried to broaden the knowledge among its members regarding international activities within the metadata area, since SCB participates in several projects. Some of these are METANET, a network for harmonising the development of statistical metadata, and METAWARE, a project for the development of a standard metadata repository for data warehouses. SCB also participates in the work of the Neuchâtel Group with the development of a conceptual model and terminology for information on classifications.

III. TODAY'S METADATA SYSTEM

7. A basic thought is that once information has entered the system, it can be used again when the same information is asked for in other parts of the system, since documentation work must not be too time-consuming or lengthy. The diagram below illustrates the metadata system of today with the arrows showing how the different parts are linked together. The broken line areas show how information in one part of the system can be copied to and directly reused in other parts.



III.1 Documentation tools/templates and their areas of use

III.1.1 Product descriptions

8. Product descriptions are requested for all official statistics according to law and are made available on the Internet. The purpose is to give brief information about the quality of the statistics and other basic facts. The template contains one section of general information and one section of a declaration of quality.

PRODUCT DESCRIPTIONS	
A. GENERAL INFORMATION	
A.1 Subject matter area A.2 Statistics area A.3 Official statistics? A.4 Responsible A.5 Producer A.6 Mandatory response?	A.7 Secrecy A.8 Destruction rules A.9 EU regulation A.10 Purpose and history A.11 Users and usages A.12 Design and implementation A.13 Planned changes
B. QUALITY DECLARATION	
B.1 CONTENTS 1.1 Statistical target characteristics • Objects and population • Variables • Statistical measures • Study domains • Reference times 1.2 Comprehensiveness	**B.2 ACCURACY** 2.1 Overall accuracy 2.2 Sources of inaccuracy • Sampling • Frame coverage • Measurement • Non-response • Data processing • Model assumptions
B.3 TIMELINESS 3.1 Frequency 3.2 Production time 3.3 Punctuality	**B.4 COHERENCE AND COMPARABILITY** 4.1 Comparability over time 4.2 Comparability between domains 4.3 Coherence with other statistics
B.5 AVAILABILITY AND CLARITY 5.1 Dissemination forms 5.2 Presentation 5.3 Documentation 5.4 Access to micro data 5.5 Information services	

III.1.2 SCBDOK

9. In 1994, the Director General decided that all observation registers and productions systems for which SCB is responsible must be documented according to a Word template with fixed headings called SCBDOK. In this case, the purpose is to provide a detailed description of how a survey is conducted from data collection to dissemination. The first chapter in SCBDOK is identical with the general information section in the Product descriptions and can be copied between the templates. A brief overview of the different sections in SCBDOK follows overleaf.

10. The first section (0) provides administrative information about the observation register. Some examples about this information are: responsible organisation and contact, purpose and history of the survey, and regulations and requirements for the European Union. As previously mentioned, this section is identical to the general information section in the Product Descriptions.

11. The second section (1) is an overview of the designing/planning phase and the contents (statistical quantities) of the survey. The output of the survey, how the statistics are published, which observation registers are created and how they are stored are also described here.

12. The data collection procedures are explained in detail in the third part (2). Here we find the frame and the frame procedure, the sampling procedure (if relevant), the measurement instrument (questionnaire with instructions on how to answer the included questions), data collection procedure and data preparation procedures (coding, editing, checking, correction etc.).

SCBDOK Version 3.0	
0 General information 0.1 Subject matter area 0.2 Statistics area 0.3 Official statistics? 0.4 Responsible 0.5 Producer 0.6 Mandatory response? 0.7 Secrecy 0.8 Destruction rules 0.9 EU regulation 0.10 Purpose and history 0.11 Users and usage 0.12 Design and implementation 0.13 Planned changes	**1 Contents overview** 1.1 Statistical observation and target characteristics 1.2 Outputs: micro data and statistics 1.3 Documentation **2 Data collection** 2.1 Frame and frame procedure 2.2 Sampling procedure 2.3 Measurement instrument 2.4 Data collection procedure 2.5 Data preparation
3 Final observation registers 3.1 Production versions 3.2 Archive versions 3.3 Experiences from latest data collection round	**4 Statistical processing and presentation** 4.1 Estimations: assumptions and formulas 4.2 Presentation and dissemination procedures
5 Data processing system	**6 Logbook**

13. Section 3 describes in detail the target and observation objects, contents in the production and archived versions, the physical organisation and storage. It is also possible to provide information on experiences from the last survey cycle.

14. Section 4 describes the estimation procedures and the presentation and dissemination procedures. The former includes assumptions underlying the estimates and computations formulas. In most cases, this information is not relevant for registers based on administrative data. The presentation and dissemination procedures can be explained in more detail here than in section 1.

15. Section 5 is a description of the data processing system. This part is for internal use only and works as a support for the producer of statistics in the day-to-day work.

16. The template finishes with a logbook containing changes implemented in the production system.

III.1.3 Metadok

17. Metadok is a software tool developed at SCB used to describe the contents of an observation register. The metadata entered into Metadok is formalised and can thus be used by other software tools. The purpose is to facilitate a multi-utilization of the metadata, e.g. for presentation on the Internet and archiving. The tool is developed in Visual Basic and the first version was released in 1998. Metadok can be used for documenting registers stored in Sybase, Access, flat files, Microsoft SQL, Paradox, SAS and Supercross.

18. Metadok is built up in a file system, where versions, databases, variables, value sets, indexes and keys for a register are described. The majority of the information is mandatory and there is on-line help for all fields to support employees in their documentation work. The first picture below shows the data model behind Metadok and the second shows the Metadok interface. Every heading in the model corresponds to a file in Metadok and every line corresponds to a field.

METADOK

1 Register

1.1 Register name
1.2 Presentation text
1.3 Description
1.4 Reference time
1.5 Register type
1.6 Subject matter area
1.7 Product
1.8 Responsible agency
1.9 Producer

2 Register version

2.1 Register version name
2.2 Presentation text
2.3 Description
2.4 Contact person

3 Database for a register version

3.1 Database name
3.2 First time
3.3 Latest time
3.4 Presentation text
3.5 Description
3.6 Database availability
3.7 Data anonymous?
3.8 Probation procedure?
3.9 Storage facility type
3.10 Storage facility id
3.11 Technical database id

4 Data matrixes in a database

4.1 Data matrix name
4.2 Presentation text
4.3 Description
4.4 First time
4.5 Latest time
4.6 Number of matrix rows

5 Variables in a data matrix

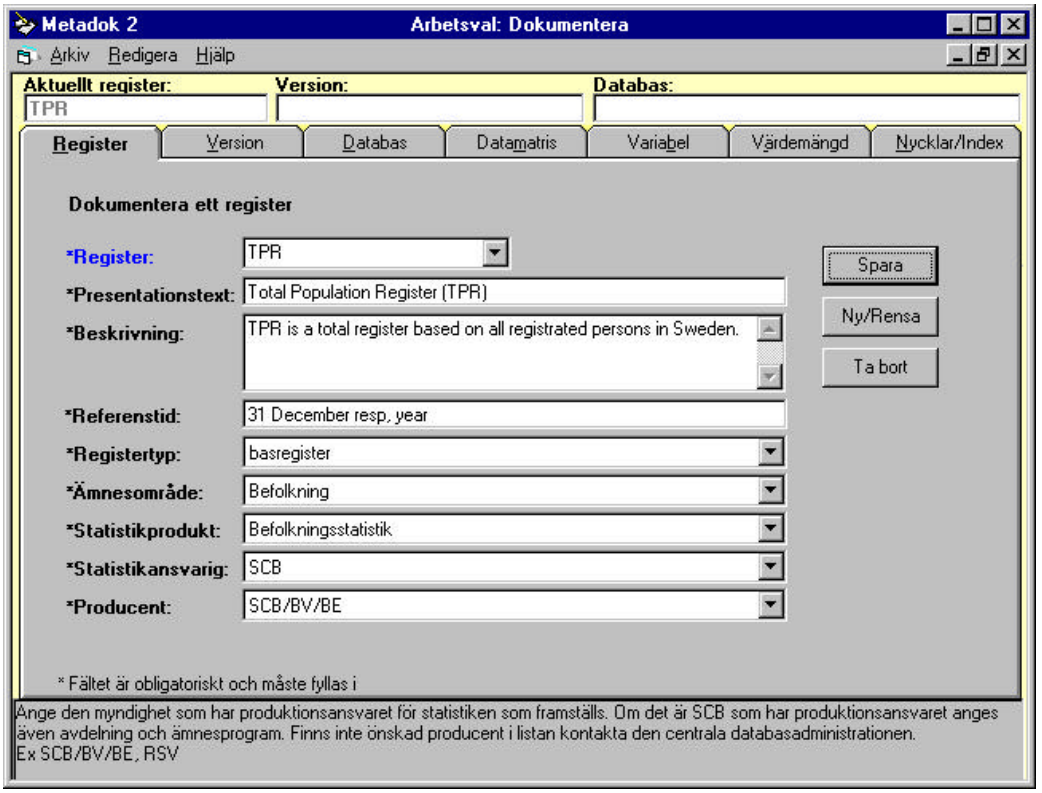
5.1 Data matrix name
5.2 Variable name
5.3 Presentation text
5.4 Variable description
5.5 Data source
5.6 Definition
5.7 Reference time
5.8 Summation?
5.9 Measurement unit
5.10 Data type
5.11 Data field start
5.12 Data field length
5.13 Number of decimals
5.14 Value set

6 Value sets for a register

6.1 Value set name
6.2 Value set description
6.3 Values exist?
6.4 Classification basis
6.5 Values (*)
6.5.1 Value code
6.5.2 Sort code
6.5.3 Value text
6.6 Variables (*)

7 Keys and indexes

7.1 Identification key (*)
7.1.1 Variable (*)
7.2 Reference key (*)
7.2.1 Variable (*)
7.3 Index (*)
7.3.1 Variable (*)



19. In order to facilitate documentation work, it is possible to import existing metadata from other sources, e.g. the Classification Database and the Sybase database where the data is stored. It is also possible to copy already documented variables and value sets from a central metadata database or local databases (authorization needed). A status control function is available for checking if the documentation is complete, since a large database can be difficult to assess.

20. The complete Metadok documentation can be exported to Word and copied into the SCBDOK template. There are also other areas of use, see below.

III.1.4 Areas of use

21. There are already a number of usage areas for documentation at SCB, and the goal is to enlarge the possibilities of benefiting from metadata in different ways. The table below shows which documentation is needed for different areas of use.

Documentation needed **Areas of use**	Product Descriptions	SCBDOK	Metadok
Intranet Presentation	X	X	X
Internet Presentation	X	X	X
Data Act Requests			X
Archiving		X	X
PC-Axis, SAS, MacroMeta			X

22. Metadok can be used in many different ways. As seen in the table, the documentation is presented on the Internet and Intranet just as SCBDOK and the product descriptions. This enables internal and external users to search among statistical products and registers, and make correct interpretations and analysis of data.

23. Since Metadok produces formalised metadata, it is possible for other software tools to use it. One example is the routine for data act requests. According to law, SCB is obliged to deliver an extract showing what data is stored at SCB about a person, if this person has submitted a data act request. Every citizen is entitled to one extract a year, which must be delivered within 30 days. The routine gathers variable and value descriptions from Metadok and these are used by the presentation of the data.

24. According to law, SCB is required to archive personal registers that are no longer used at the National Archives. This also concerns non-personal registers considered to be of future interest. Irrespective of external demands, SCB often has an interest in archiving its registers. Former years' issues of registers are frequently used in work on commission for research purposes. Because of archiving, the Metadok documentation is used in an automatic archiving routine. This process uses register data from the production system and converts these to the approved format. When needed, the process can also return registers into the production environment. In addition to this, a complete SCBDOK documentation has to be sent to the archive.

25. PC-Axis for micro data is an aggregation tool where metadata from Metadok is utilized. Variable- and value labels entered into Metadok are used in PC-Axis by the creation of tables directly from micro registers. Values defined in Metadok can also be exported into SAS and be used in statistical processing. MacroMeta is a tool especially developed for entering metadata into Sweden's Statistical Databases (see below) and this tool can also use values described in Metadok.

III.2 Other parts in the metadata system

26. The documentation tools/templates and their areas of use are not the only parts of the metadata system. The classification database and metadata for aggregated data are also considered important parts of the system.

III.2.1 The Classification Database

27. The classification database is an important coordination instrument at SCB containing both national and international classifications, e.g. regional and industrial groupings, and keys between different classifications. The software tool used for the classification database is Bridge. At the time of writing, the work with making the database available for both external and internal users is underway.

III.2.2 Metadata for aggregated data

28. Another part in the metadata system is metadata for multi-dimensional tables in Sweden's Statistical Databases (SSD). The first version of SSD was launched on the Internet in 1997 and contains statistics (macro data) for around 20 different subject areas. In order to create these tables, metadata is stored in a central database. The subject area departments are responsible for this work as well.

29. The data model for SSD is very complex and will thus not be discussed here. Those who want to know more about SSD can visit SCB's website, which is available in English on www.scb.se/eng, under Sweden's Statistical Databases. For the time being, there are two subject areas translated into English, population and prices and consumption. More detailed information can be found in Sweden's Statistical Databases: Detailed Descriptions of the Metadata in the Macro database version 1.02 (Iréne von Rothstein & Bo Sundgren, 1996).

IV. CONCLUSIONS AND FUTURE PLANS

30. As mentioned in section II, the Metadata project has been striving to obtain a basis for future development of the metadata system. Our conclusion is that three main factors are of decisive importance for documentation and the metadata system: technology, knowledge and attitude.

IV.1 Technology

31. In order to make the system more functional and user-friendly, further deployment of the already existing parts and connections as well as new applications is necessary. It is important that information that has once entered the metadata system can be reused when the same information is asked for in other parts of the system. To a great extent, this is a question of technology since one crucial thing is to create efficient links between the different parts of the system.

32. As previously mentioned, an extensive mapping of the needs and wishes concerning the metadata system has been done. Many valuable viewpoints have come to light in connection with the documentation courses and at meetings with representatives for the subject area departments. In the short term, there is a demand for technical development of Metadok. Apart from a number of minor corrections and improvements, a link between Metadok and SAS/Power designer will be implemented. It is essential that the existing tools and links work satisfactorily; otherwise many will await further development instead of starting to work with their documentation right away. In the long run, many voices have been raised for the creation of a possibility to use metadata during the whole production process. This is not sufficiently supported by the system today and an investigation of how this can be done must be carried out.

33. Today, there is a wide variety of variables in the registers and surveys at SCB and there is a strong need for coordination between different subject areas. The lack of documentation has complicated

comparisons of variable definitions and descriptions. Examples show that there are variables with the same meaning but with different names in different surveys. Likewise, there are variables with the same name but meaning different things. A precondition for an efficient production of statistics is that the registers and surveys at SCB are coordinated in terms of technology and contents. We think that a way to solve this problem is to build a variable database, which can serve as a coordination instrument. Exceptional solutions, which otherwise can cause disorder in the interpretation and analysis of data, can then be avoided. At the time of writing, no long-term decisions have been made, but a discussion about the conceivable functionalities is underway.

34. There will also be an evaluation of Bridge, the tool used for the Classification Database, in order to decide whether this can be used in other parts of the metadata system, e.g. for a variable database since this has been brought forth as an important future development. This evaluation will be completed in February 2002.

IV.2 Knowledge and attitude

35. It can be established that there is a great need for education in the metadata area regarding the available tools and templates, how the parts in the metadata system are connected, and about the usage areas for completed documentation.

36. After this year's documentation courses, we have noticed an increased documentation activity, but it is obvious that some support is still needed. In connection with the course evaluation, the most important obstacle for completing documentation with Metadok was asked for. The majority states that lack of time is the main problem, but more education, support and possibilities to use metadata in the production process are also considered important. Since the courses have been much appreciated and there is still a demand for more support and education, the concentration on documentation education will continue in 2002.

37. Experience shows that knowledge about the metadata system and its performance is not always the problem; many times there is a lack of understanding for how data is stored. In order to carry out a complete documentation, there is a need for knowledge about the subject area as well as about IT matters, and thus cooperation between these categories is necessary. Working with documentation and metadata is traditionally considered as boring and has a low status. Because of this, it is given low priority and turns out to be one of the last duties to be carried out. Since there is usually more work to do than there is time available, this area suffers. For example, we have noticed that the newly employed are often are given the documentation task since others in the subject-matter programs lack time and/or inclination. It is a good way to learn, but strong support from someone more experienced is crucial. In many cases this kind of support is missing and there is nothing written down about the product/register in question.

38. Changing attitudes is a time-consuming task but it is necessary to influence the attitudes towards documentation in order to attain the goals that are set within the metadata area. It is important to make the benefits of metadata visible in order to motivate those working with documentation. If this is not achieved, we could have the perfect system from a technical point of view and might still be struggling with the same lack of metadata as today.

REFERENCES

Bo Sundgren, The Swedish Statistical Metadata System, SCB, 19 January 2000.

Iréne von Rothstein & Bo Sundgren, Sweden's Statistical Databases: Detailed Descriptions of the Metadata in the Macro database version 1.02, 1996.