THE METADATA SYSTEM AT STATISTICS SWEDEN IN AN INTERNATIONAL PERSPECTIVE

Hans Lindblom, Deputy Director General, and Bo Sundgren, Senior Adviser,
Statistics Sweden

Summary

The metadata infrastructure of a statistical office should provide different kinds of metadata, serving the needs of different kinds of users. The metadata systems need to be organized in such a way that metadata can be captured and processed in an efficient way, without too much manual work and with efficient computer support. The paper describes on-going efforts at Statistics Sweden to achieve these goals, including both problems and opportunities that have been encountered. One important conclusion from this work, and from discussions with other statistical agencies, is that metadata systems in national and international agencies need to be better harmonized in order to facilitate international exchange of statistical data and metadata. The paper presents some ideas about how this could be achieved in the future.

1 WHAT IS "STATISTICAL METADATA", WHY ARE THEY NEEDED, AND HOW CAN THEY BE OBTAINED?

Metadata are data about data, data informing about different aspects of data: contents-oriented aspects, technical aspects, and others. The concept of metadata is usually interpreted so as to include not only metadata that directly and explicitly describe data, but also metadata that describe the processes behind the data, as well as the resources needed by these processes; the last category of metadata includes metadata that may also be classified as administrative data, e.g. data about costs and revenues.

Statistical metadata are data about statistical data. Statistical metadata are needed for a number of purposes, oriented towards the needs of users and producers of statistical data, respondents, managers, and funders.

The main purpose of statistical metadata is to inform users of statistical data about

- which statistical data are available
- where to find and how to retrieve certain statistical data that they need
- how to interpret statistical data, once they are available

Another major purpose of statistical metadata is to promote the quality and efficiency of statistical production processes. By using metadata actively for "driving" production processes, these processes are controlled in a more automatic and flexible way; it will not be possible to execute such a process, unless appropriate metadata are available, and on the other hand, if the input data and metadata are for some reason changed (within permitted limits), the processes will adapt automatically to these changes. By designing statistical production processes so as to produce metadata about their own performance, so-called process data, malfunctions and inefficiencies may be detected promptly, and may be corrected by managers of the processes, or sometimes even automatically.

Designers of statistical systems are interested to obtain metadata about similar systems, from their own office or from other offices, when they design, construct, and implement a new system. For already existing systems under their responsibility they are interested to obtain feedback about the performance (qualities and costs), usage, and user satisfaction.

On a higher level, managers and funders of statistical production processes will be interested to learn if the users of the statistical data produced are getting value for the money: to what extent do users actually use the statistical data, and are they satisfied with the qualities of the data: contents, accuracy, timeliness, availability, comparability, coherence, etc?

Even respondents can benefit from good metadata. Metadata can be used for explaining the purposes of a statistical survey, as well as for giving instructions about how to complete questionnaires and provide data.

A key issue in connection with metadata is how to obtain them. Producing metadata as a separate activity in the statistical production process is resource consuming. A more efficient approach is to capture metadata from other sources wherever possible, and to design production processes so as to produce data and metadata in parallel, and in such a way that data from one process may, possibly after some automatically or computer-supported transformations, be used by other processes. In particular, it should be recognized that statistical design processes typically result in design decision that explicitly or implicitly contain metadata about the system under design and its components, processes, and resources; these metadata should be systematically, and preferably automatically, captured by documentation and metadata systems operating together with the design processes in a well integrated way. Also the operation processes, as has already been mentioned, can be designed so as to produce useful metadata, so-called process data.

2 METADATA SYSTEMS AT STATISTICS SWEDEN – THE CURRENT SITUATION

Figure 1 illustrates what is sometimes called the statistical data/metadata life cycle or value chain; cf Porter (1985), Sundgren (2003a), ECB&Eurostat (2003)

During the life cycle statistical data and accompanying metadata pass through four relatively well-defined stages, corresponding to forms and interfaces:

Stage 1: The input data stage: the input data/metadata as registered on some kind of input form, e.g. a completed (paper or electronic) questionnaire.

Stage 2: The final microdata stage: the input data/metadata as finally stored in some kind of final observation register, e.g. a relational database, after data preparation operations such as coding, editing, and other transformations (e.g. computation of derived variables).

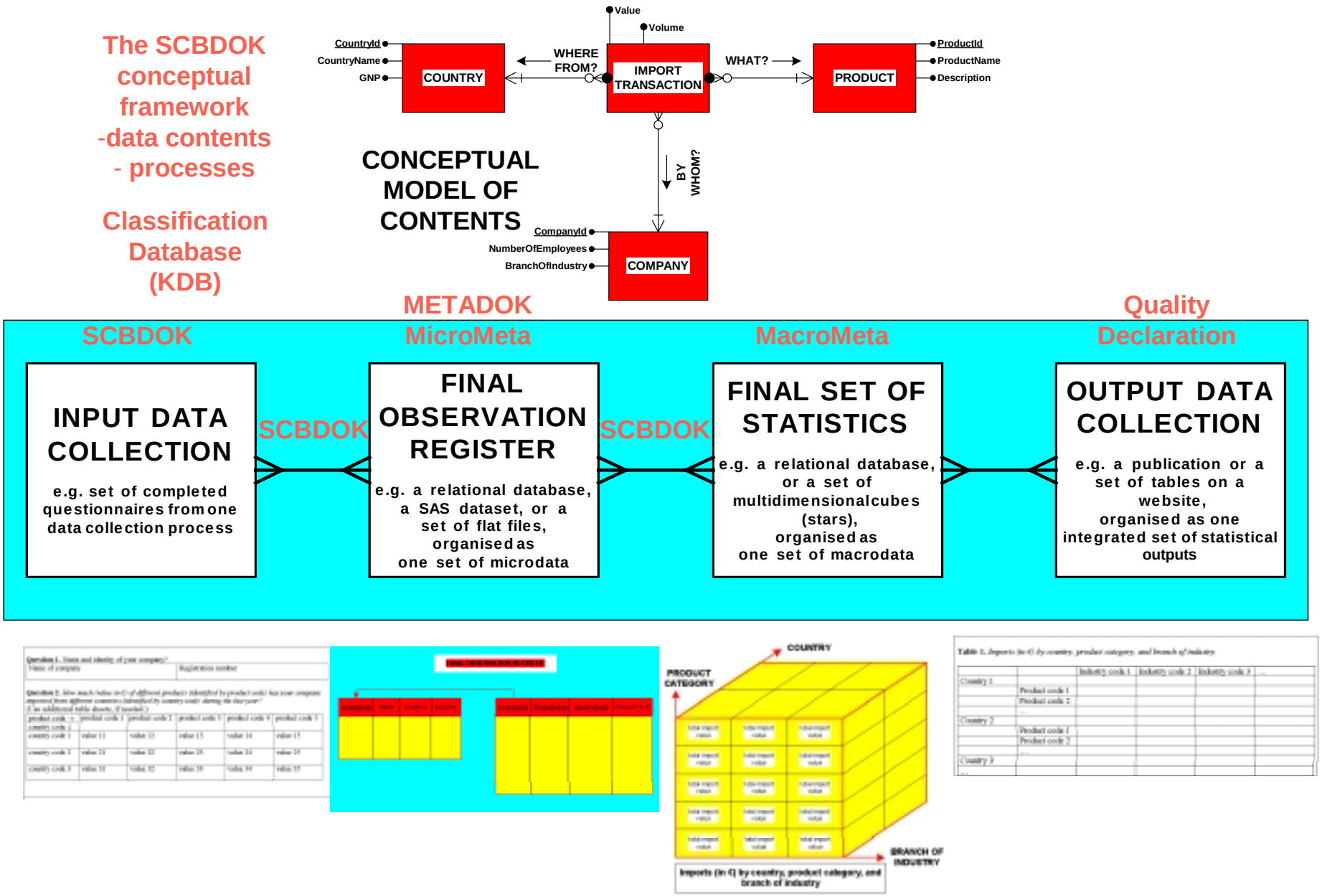

Figure 1. Fundamental data/metadata interfaces and metadata objects in a statistical production system.

Stage 3: The final macrodata stage: the output statistics (estimated values of statistical characteristics) and accompanying metadata as finally computed and stored in some kind of output database.

Stage 4: The output product stage: the statistical data/metadata as published and disseminated via printed and electronic media.

As indicated by the "fork arrows" (⋟⋞) between the four boxes in figure 1, there are "many-to-many"-relationships between the four stages, i.e. the same input observations used in several observation registers, each one of which may be used in the production of several output data sets, which again may be combined into several statistical end-products; and vice versa: a certain statistical end-product may be based upon several output data sets, each one of which may be derived from several observation registers, which again may be the result of combining input data from several sources. It should be noted that these complex relationships between inputs, throughputs, and outputs already exist to a high degree in modern statistics production, although most statistical offices are still organized according to the traditional stovepipe model; the production system logic is not necessarily isomorphic with the organizational structure. This is something that has to be carefully considered when designing statistical metadata systems.

1.1 The SCBDOK conceptual framework

All metadata systems and documentation templates at Statistics Sweden are based upon the same conceptual framework – the SCBDOK framework, first described in Rosén&Sundgren (1991). According to this framework the real world is conceptualized in terms of

- objects, e.g. persons, organizations, events; a population is a set of objects
- relations between objects, e.g. the employment relation between persons and organizations: employment (person, organization)
- variables of objects, e.g. age (person), size (organization), time (event)
- values of variables; e.g. "25 years" could be the value of the age of a person; the values come from value sets, sometimes referred to as classifications

Statistical microdata (observation data) represent measurements of values of variables of individual objects in a population. Statistical macrodata (aggregated data, statistics) represent estimated values of parameters of populations. A parameter of a population summarizes the (true) values of a variable of the objects in the population by means of some kind of summation function, called statistical measure, e.g. frequency counting, arithmetic summation, mean computation. "Average age" could be a parameter of a population of persons, a population of buildings, or a population of cars.

By applying a statistical measure, m, on the (true) values of a variable, V, for the objects in a population, O, we obtain the (true) value of a statistical characteristic, O.V.m, where V.m is a parameter, and O.V is sometimes called an object characteristic.

An estimated value of a statistical characteristic, O.V.m, is obtained by applying an estimation function, called estimator, e, on a set of measured values of the variable V of observed objects in the population O.

Because of different kinds of errors and uncertainties there will always be a discrepancy between the estimated and the true value of a statistical characteristic. One important purpose of statistical metadata is to describe and, if possible, estimate the size of these errors and uncertainties, e.g. coverage errors, sampling errors, response errors, measurement errors, processing errors. The sources of the errors and uncertainties are to be found in the design and operation processes of a statistical system (cf figures 2 and 3). In order to describe the errors properly, it will usually be necessary to describe the processes behind in some detail. Thus the SCBDOK framework contains a process model as well as data models; cf figure 2.

Another important purpose of statistical metadata is to describe the contents and meaning of statistical data, so that a user can judge the relevance of the data for his or her information needs. The sources of these metadata are to be found in the design process and in particular in the resulting conceptual model of the contents of a statistical system in terms of the concepts briefly described above: statistical characteristics, populations, objects, relations, variables, value sets (classifications), etc.

The contents descriptions and the error descriptions together constitute important parts of a quality declaration of statistical data.

Figure 4 gives an overview of some important concepts in the SCBDOK framework. It focuses on the distinctions between

- the (true) reality itself and the reality as observed and estimated by statistical data
- statistical characteristics and object characteristics (on the reality level)
- macrodata and microdata (on the observation and estimation level)

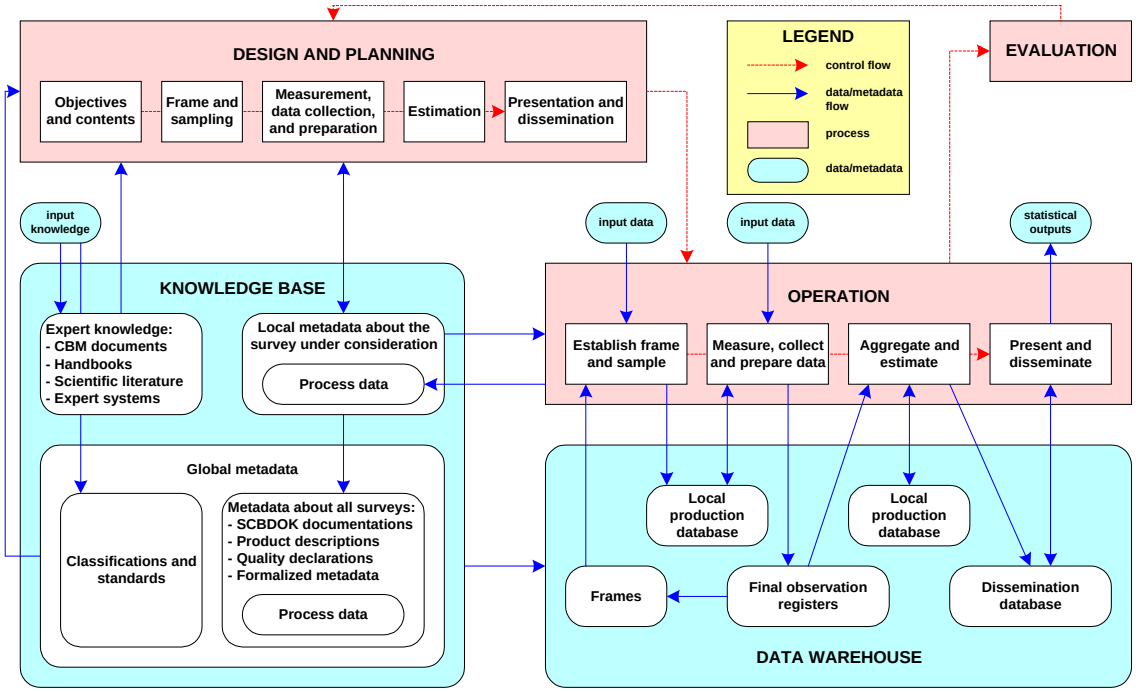

Figure 2. Processes and data/metadata sets in statistics production.

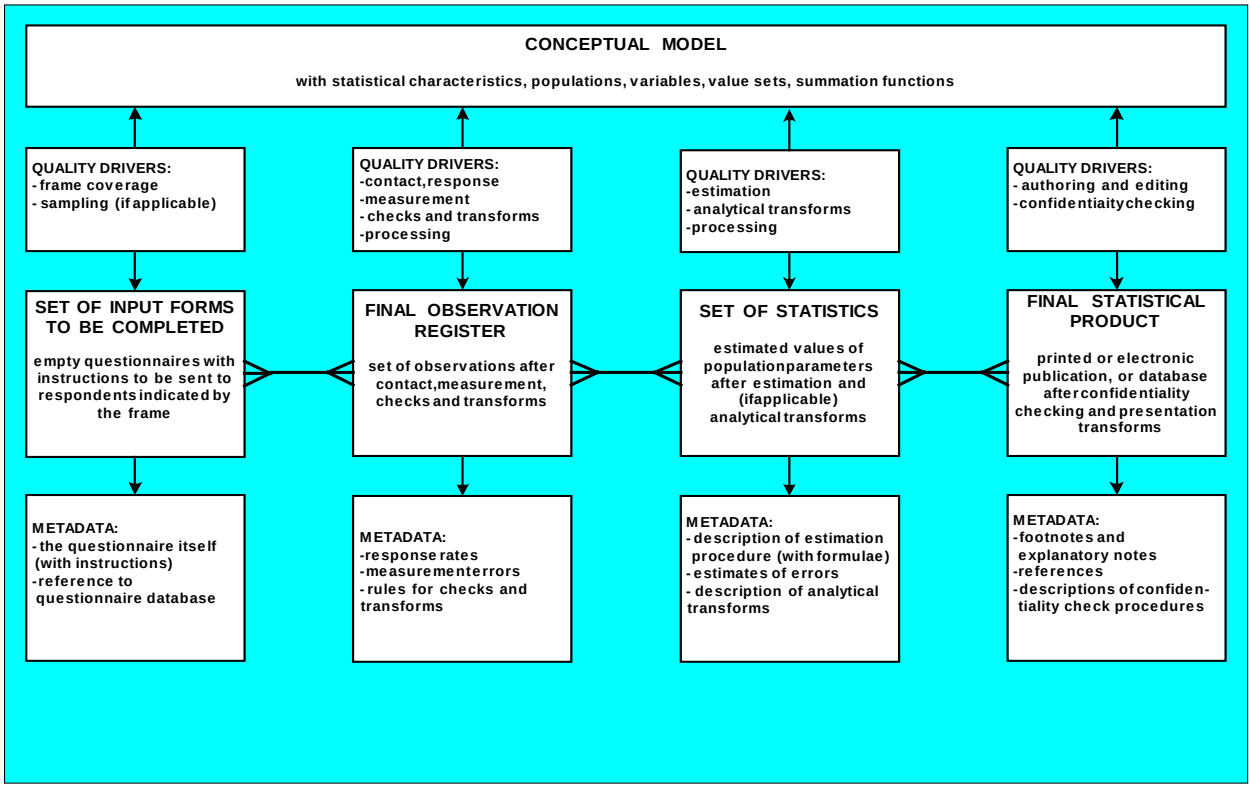

Figure 3. Conceptual model, quality drivers, and required metadata in statistics production.

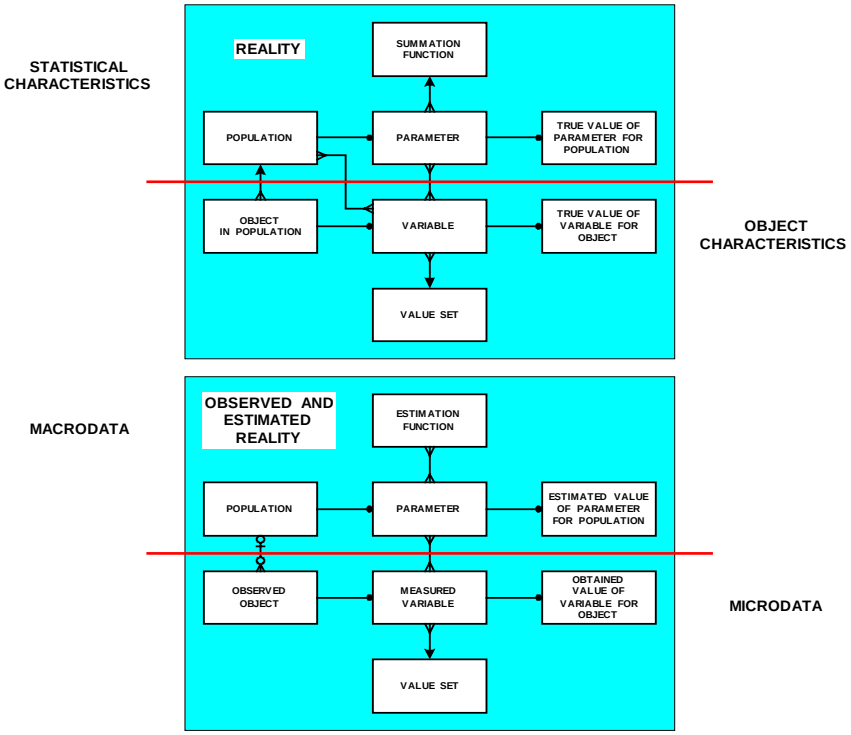

Figure 4. Fundamental concepts in statistics production.

1.2 The SCBDOK documentation template

Figure 5 gives an overview of the SCBDOK documentation template, version 3.0.

SCBDOK 3.0	
0 General information 0.1 Subject matter area 0.2 Statistics area 0.3 Official statistics? 0.4 Responsibility 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 General approach to implementation 0.13 Planned changes	**1 Contents overview** 1.1 Observation characteristics 1.2 Statistical target characteristics 1.3 Outputs: microdata and statistics 1.4 Documentation and metadata **2 Data collection** 2.1 Frame and frame procedure 2.2 Sampling procedure (if applicable) 2.3 Measurement instruments 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 the latest 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**

Figure 5. The SCBDOK documentation template, version 3.0.

The major contents of SCBDOK documentation could be described as follows (cf also figures 2-4 above):

Chapter 0 contains administrative information and is aimed at managers and others who need rather superficial information about the survey. Chapter 0 is also used separately from the complete SCBDOK documentation, and it may then be called “product overview”.

Chapter 1 gives an overview of the contents of the documented statistical system[1] in a structured way. The observed (or derived) object characteristics and their relations to each other are specified by means of an object graph (such as the one on the top of figure 1), and the statistical target characteristics are specified in terms of population, classification variables, summation variables, and statistical measures. The outputs are specified, both microdata (final observation registers) and macrodata (statistics). Finally, references are made to other relevant sources of documentation and metadata, e.g. methodological reports.

Chapter 2 gives a detailed description of the data collection process, including frame and frame procedure, sampling procedure (if applicable), measurement instruments (typically questionnaires in some form or other), the data collection procedure proper, and data preparation procedures (data entry, coding, editing and correction, imputation, production of

[1] The term “survey” is often used instead of “statistical system”, but the focus of SCBDOK is on all kinds of statistical systems, not only statistical surveys in the traditional sense, but also statistical systems based on administrative data, registers, and secondary systems like indexes and national accounts. Thus if the term “survey” is used, it should be interpreted in a generic sense.

derived objects and variables). The detailed metadata should include all rules, instructions, and practices that have an impact on the meaning and quality of the data.

Chapter 3 describes both the contents and the storage and other technical aspects of the final observation registers. There are often several versions of the final observation registers, possibly with different contents and managed by different software. It is important to distinguish between production versions, e.g. versions managed by a database management system, and those versions that are submitted to an archive for reuse in the near or (very) distant future. The latter versions of the final observation registers must be stored in such a way that they will last for a long time and can be reused without access to the hardware and software that we use today.[2]

Chapter 3 also contains an item for so-called process data, that is, metadata that describe circumstances and events that are unique for each repetition of the survey and its processes, e.g. data about response and non-response.

Major parts of chapter 3 are supported by a software tool called METADOK (cf figures 6 and 7 below) that ensures that the metadata that are captured in this part of the documentation are structured and formalised in such a way that they can easily be used by software used in statistics production. Among other things this facilitates the implementation of metadata-driven statistics production systems.

Chapter 4 describes the estimation procedures, including mathematical formulas and assumptions made, as well as presentation and dissemination procedures.

Chapters 1, 2, and 4 contain the documentation basis needed for the production of so-called quality declarations that (together with the product overviews mentioned above; chapter 0 in SCBDOK) are mandatory for published official statistics in Sweden, regardless of whether they are published electronically or by means of traditional paper publications. Figure 8 shows the quality declaration template.

A handbook is available in Swedish providing detailed instructions and examples of how to complete the SCBDOK template, item by item – for each item there is a checklist of subitems to be covered. Cf Statistics Sweden (2001b).

2.3 MicroMeta: metadata model for final observation registers

Figure 6 gives an overview of the metadata model used by the METADOK documentation tool when describing final observation registers. The complete model is illustrated by figure 7. A detailed description of this model is given in Statistics Sweden (2003). The final observation registers described may be stored as, for example, relational databases, SAS data sets, or flat files. The model corresponds to section 3 of the SCBDOK documentation template, version 3.0.

[2] Some statistical systems, e.g. the national accounts, regularly produce a number of clearly distinguished versions like “preliminary data”, “final data”, and “revised data”. All such versions are associated with their own “final observation registers”; thus even the preliminary data will reach the status of a final observation register at some well-defined point in the production process.

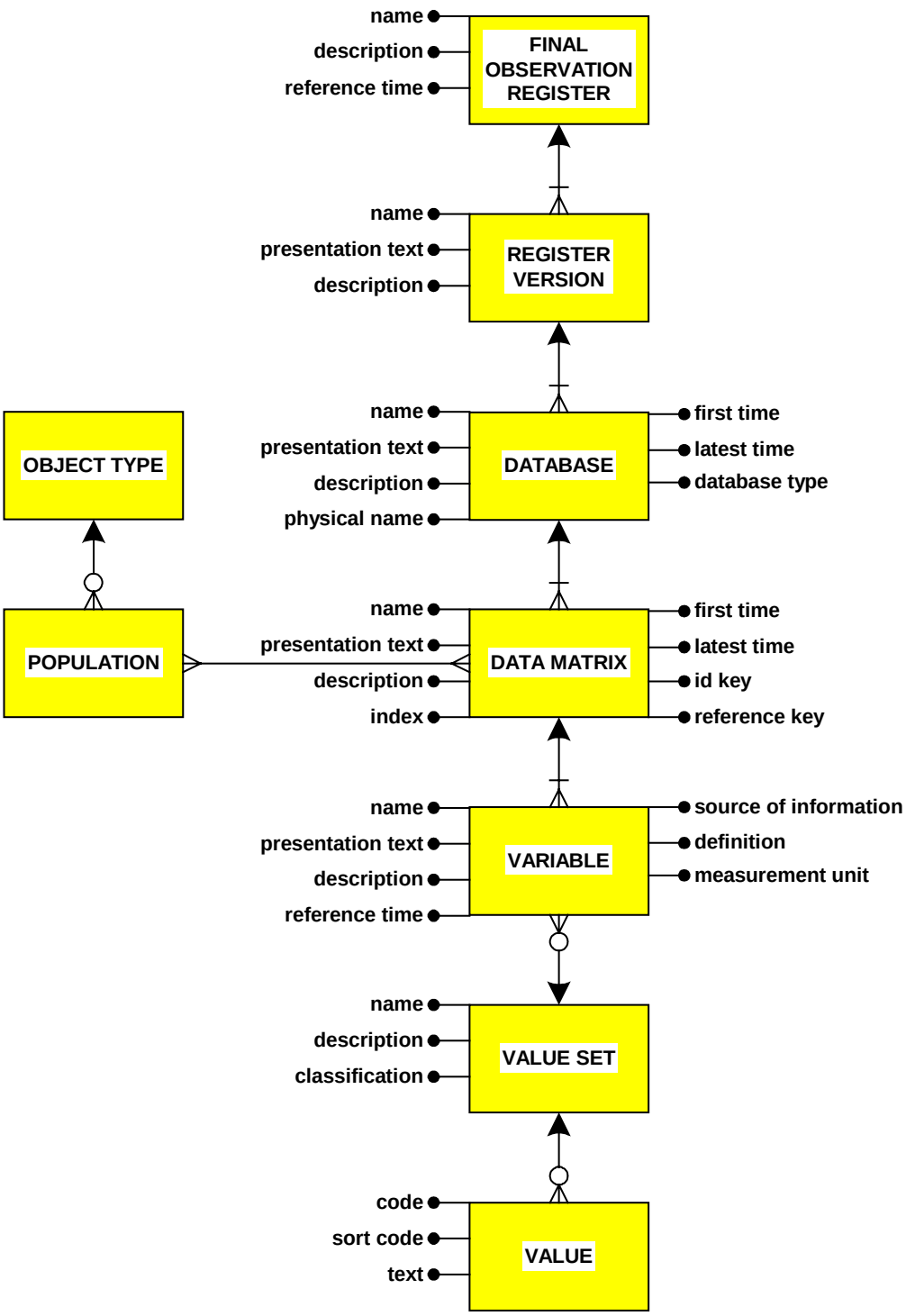

Figure 6. Conceptual model of basic metadata for a final observation register.

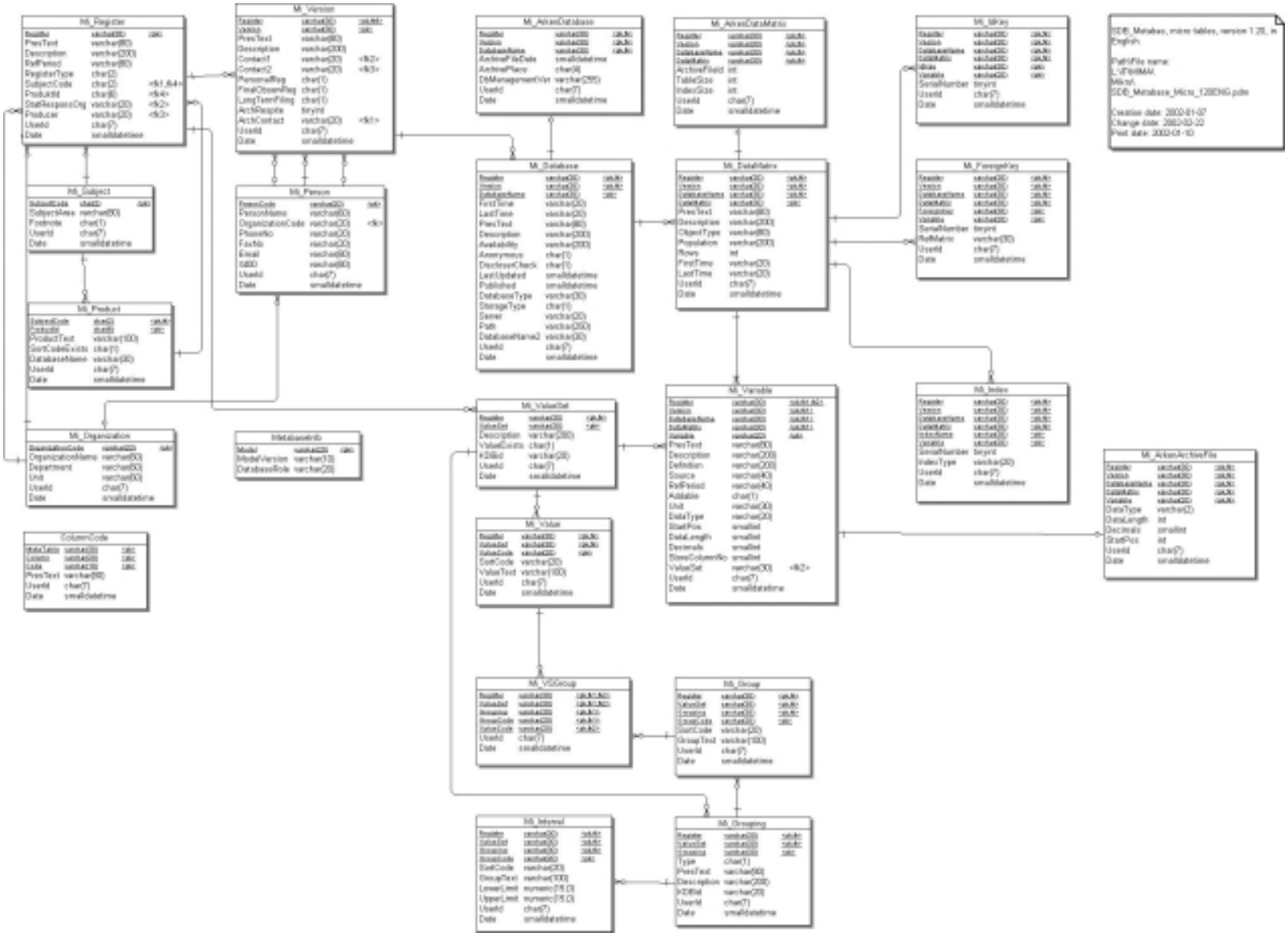

Figure 7. The MicroMeta data model

2.4 The Quality Declaration Template

Figure 8 gives an overview of the Quality Declaration Template used by Statistics Sweden.

Quality Declaration Template	
1 Contents 1.1 Statistical target characteristics 1.1.1 Objects[3] and population 1.1.2 Variables 1.1.3 Statistical measures 1.1.4 Study domains 1.1.5 Reference time 1.2 Comprehensiveness	**2 Accuracy** 2.1 Overall accuracy 2.2 Sources of inaccuracy[4] 2.2.1 Sampling 2.2.2 Coverage 2.2.3 Measurement 2.2.4 Non-response 2.2.5 Data processing[5] 2.2.6 Model assumptions 2.3 Presentation of accuracy measures
3 Timeliness 3.1 Frequency 3.2 Production time 3.3 Punctuality	**4 Coherence especially comparability** 4.1 Comparability over time 4.2 Comparability over space 4.3 Coherence in general
5 Availability and clarity 5.1 Forms of dissemination 5.2 Presentation 5.3 Documentation 5.4 Access to microdata 5.5 Information services	

Figure 8. The quality declaration template used in the Swedish statistical system.

The Quality Declaration Template, and the concepts used there, is discussed in detail in Statistics Sweden (2001a). The conceptual foundation of the Quality Declaration Template is the same as that of SCBDOK and is very close to the model used by the so-called European Quality Concept used in (slightly different versions) by Eurostat, OECD, and most member states of the European Union; cf Eurostat (2003a, 2003b) and OECD (2003).

Figure 9 illustrates the origins and effects of different types of errors in a statistical production process. The mechanisms behind these errors have to be described in detail in SCBDOK and summarised in the Quality Declaration for the needs of end-users of statistics (macrodata).

[3] Actually the present version of the Quality Declaration Template uses the term ”unit” here. In order not to confuse the reader of this paper and many other papers concerning SCBDOK, we will use the standard term “object” here. In SCBDOK the term “unit” stands for “measurement unit”, e.g. US dollars, tons.
[4] Also called ”sources of error”.
[5] The term used elsewhere in this paper and in other papers concerning SCBDOK is ”data preparation”.

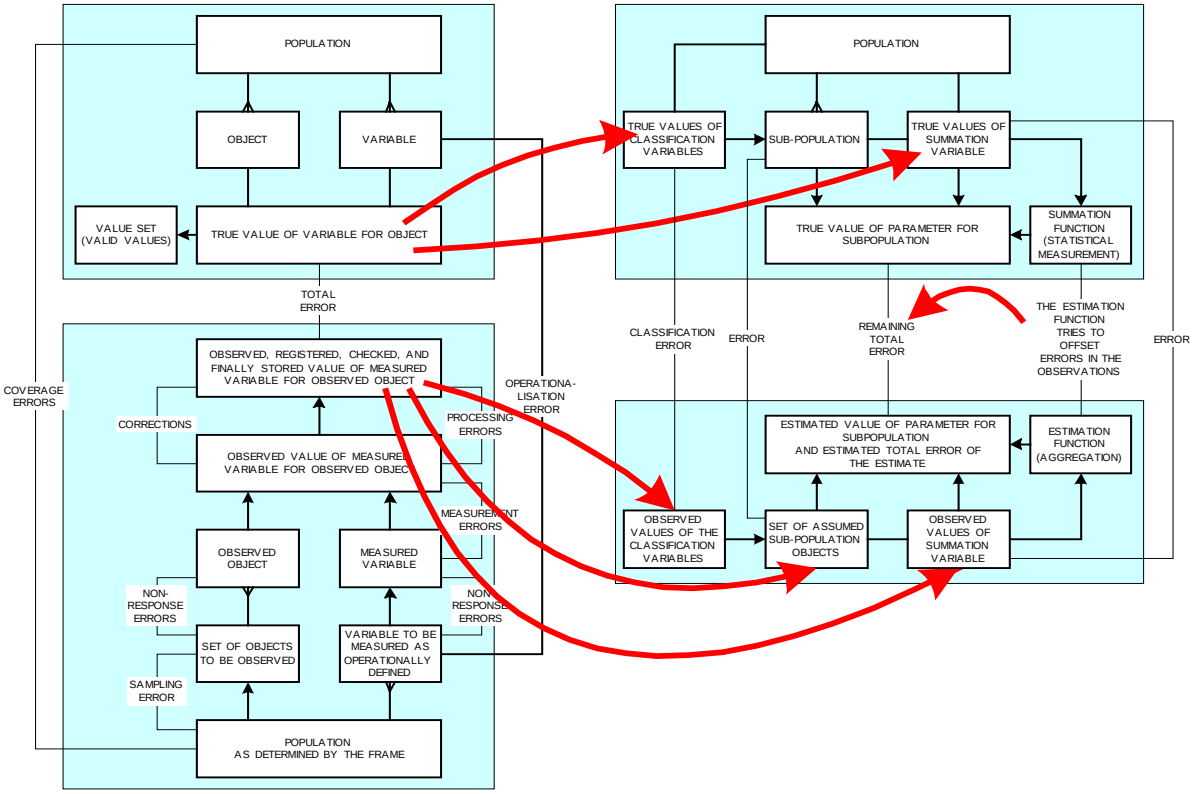

Figure 9. Origins and effects of errors in statistics production.

2.5 The conceptual model underlying the Classification Database (KDB)

The model is identical with the Neuchâtel model, version 2.0; Neuchâtel Group (2002).

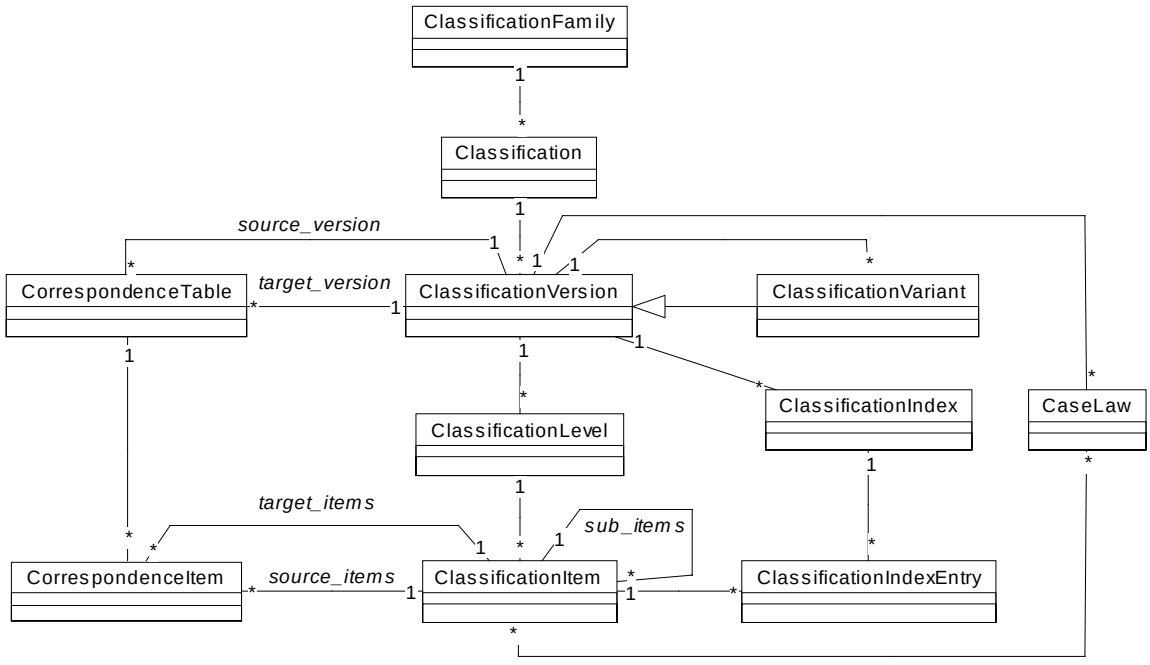

Figure 10. Conceptual model for Classification Database (KDB) at Statistics Sweden.

2.6 MacroMeta: metadata model for macrodata used by the Statistics Sweden's output database

Figure 11 gives a simplified overview of the metadata model used by the software system supporting the dissemination or output databases of Statistics Sweden, which are available on the Internet at www.scb.se . The complete MacroMeta model, containing among other things support for footnotes and bilingualism, is shown in figure 12.

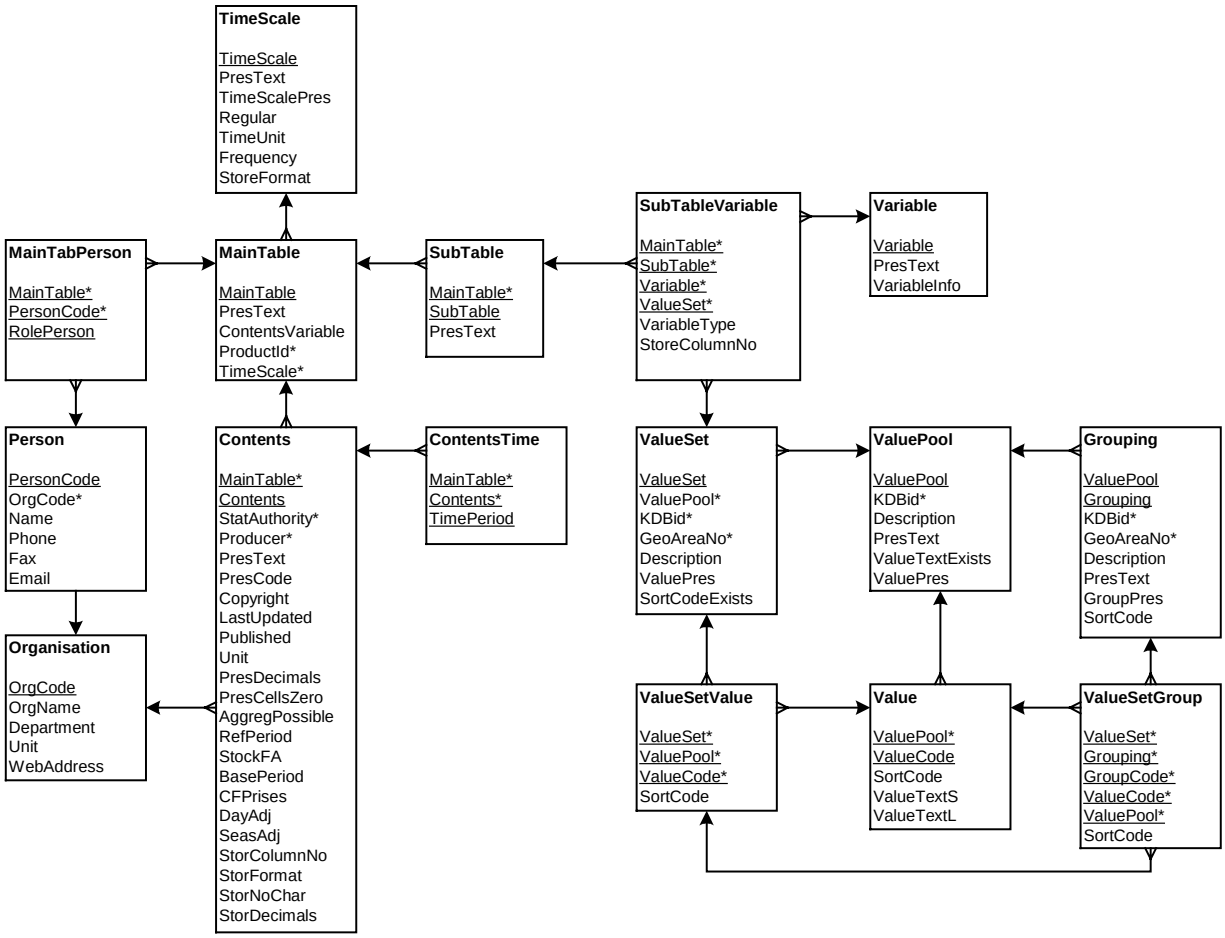

Figure 11. Relational data model for basic metadata for final statistics stored in a relational database.

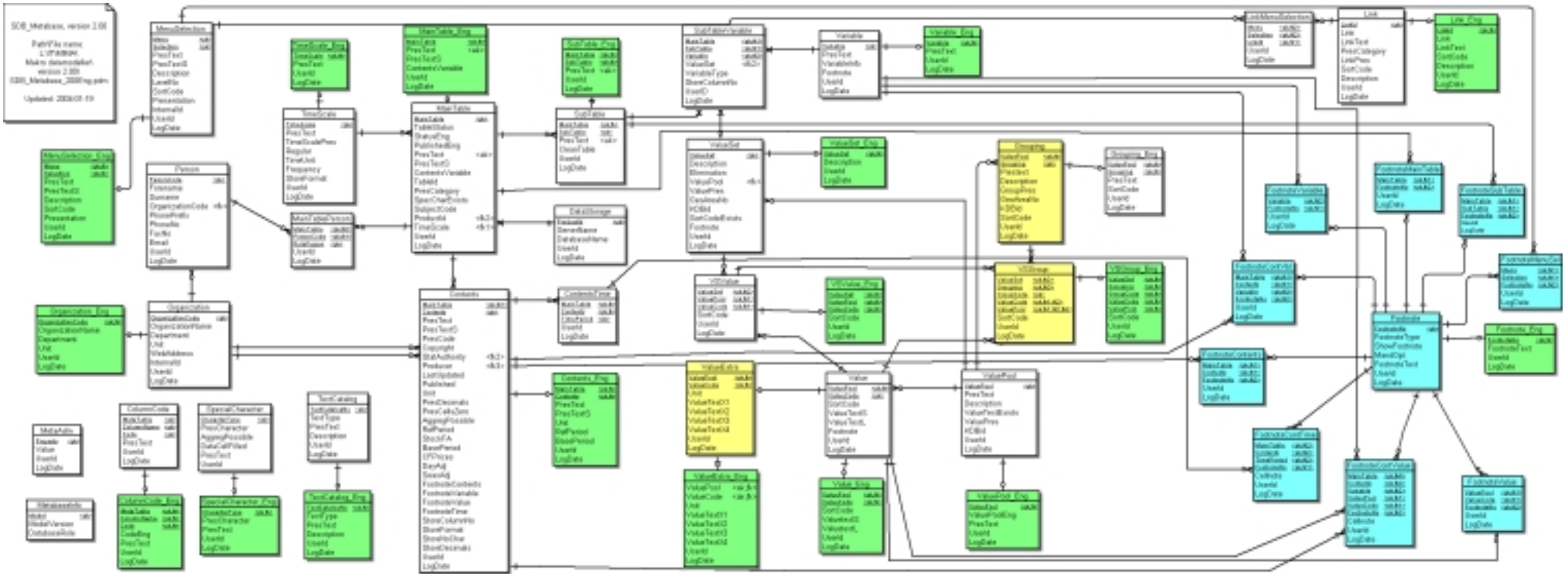

Figure 12. Relational data model for basic metadata for final statistics stored in a relational database.

3. THE HISTORY OF METADATA AT STATISTICS SWEDEN

Statistics Sweden has struggled with metadata for more than three decades. Already in the early 1960's, the Norwegian Svein Nordbotten launched his vision of a statistical data warehouse[6], based on standardized microdata files, systematically documented in a data catalogue[7], and managed by standardized processes supported by generalized software; Nordbotten (1960, 1966). The top management of Statistics Sweden became interested in Nordbotten's ideas, and started a number of development projects at the end of the 1960's with the intention to reengineer the production processes of Statistics Sweden from both a technical and an organizational point of view. The data warehouse would include both microdata and macrodata in standardized form, and the data would be described in a catalogue of variables, both from a technical and from a contents-oriented perspective. Microdata and macrodata processes would be driven by standardized software.

The privacy debate triggered by the 1970 population census in Sweden made it impossible for Statistics Sweden to continue the development of a data warehouse including microdata. The contents of the data warehouse had to be limited to aggregated statistics on a relatively high level. In 1976 Statistics Sweden launched its first online database, available to external users, and including a wide range of statistics, e.g. socio-demographic statistics, economic time series, and regional statistics. All data were managed by the AXIS database management system, developed by Statistics Sweden. The system was metadata-driven, and the metadata model used for that system is still used, in modified form (the MacroMeta model, see chapter 2 of this paper), by the current Internet-based output databases, Sweden's Statistical Databases, which were launched by Statistics Sweden in 1996.

The development work at Statistics Sweden in the early 1970's also resulted in a generalized, metadata-driven software product for tabulations, TAB68, which could easily be used even by non-programmers. Later developments resulted in a whole family of generalized software products, based on the TAB68 program code, for statistical processes like data editing, file matching, data transformations, and variance computations. A more recent development in the same tradition is the PC-AXIS software for user-friendly retrieval and manipulation of statistical data. PC-AXIS is also based on the metadata models described in this paper.

The theoretical basis for these developments can be studied in Sundgren (1973), where the term "metadata" is first used. This thesis is based on the already mentioned work between Nordbotten, and on the seminal book by Langefors (1966), where the distinction by information and data is made clear, and where a comprehensive theory of information systems is presented for the first time. Sundgren (1973) formulates an infological theory of databases and introduces conceptual modeling as a systematic way of describing the contents of databases and information systems. This conceptual framework is further developed for statistical purposes in Rosén&Sundgren (1991).

The ambitions to introduce metadata in the statistical production processes in a more systematic way were successful in the sense that some of the generalized software products developed (TAB68, AXIS, PC-AXIS, Sweden's Statistical Databases) were clearly metadata-driven, based on explicitly defined metadata models. However, the metadata ambitions utterly failed in another sense. As early as 1971 a project was started to develop a so-called catalogue of variables. This metadata system was supposed to contain complete and detailed descriptions

[6] The term used by Nordbotten was "statistical file system" or "archive-statistical system".
[7] The term "metadata" was introduced by Sundgren (1973).

of all surveys and data sets maintained by Statistics Sweden, and it was aimed to cover both technical and contents-oriented aspects. This project was initiated by the top management of Statistics Sweden, but unfortunately the top management was not able to convince the whole organization of the benefits of such a system, and some parts of the organization passively or actively resisted the efforts. Even those parts of the organization, which made their part of the work, had to give up after some time, since the system turned out to be of limited use to them, if it did not contain metadata from other parts of the organization. This demonstrates the importance of viewing metadata systems as corporate assets, parts of the infrastructure of the organization. The experiences of this project also demonstrate that many professionals regard the knowledge (metainformation) that they possess as a personal asset rather than as something that belongs to the organization as a whole, and which should be actively and systematically shared with others. These experiences are in line with more recent research findings in the area of knowledge management: knowledge monopolies are sometimes hard to break, and knowledge-based organizations are more dependent on individual persons than they would like to admit. Statistical offices are sometimes described as rather loose associations of small kingdoms.

In 1974 the top management of Statistics Sweden presented a detailed plan to restructure the organization of Statistics Sweden from a traditional stovepipe organization to an organization focused on a data warehouse with input processes feeding data from surveys and administrative sources into the warehouse, and with separate output processes combining data from the data warehouse into statistical end-products tailored to the needs and requirements of different kinds of users. This proposal created such opposition from middle management and from the trade unions that it had to be withdrawn. The stovepipe organization still prevails at Statistics Sweden, even though the production processes nowadays form much more complex input-throughput-output patterns, as was already mentioned in connection with figure 1.

4 STRENGTHS AND WEAKNESSES

The history of metadata at Statistics Sweden, briefly summarized above, contains both successes and failures. The first attempt to introduce a corporate metadata system, the catalogue of variables, was clearly a disaster. The only positive thing about such disasters is that you can actually learn a lot from them – if you are mentally prepared for it. Some of the lessons that Statistics Sweden has learnt from this project are reflected in the Golden Rules presented in the next section (figure 13).

The successes have been slower and less dramatic. Very often you have the feeling that there are only problems associated with metadata systems: very much talk and very little action. However, when you write a paper like this one, you realize that there is some progress; even if it is much slower than we would like it to be.

We have already mentioned the metadata models and the metadata-driven software products developed since the 1970's, and which have resulted in the MicroMeta and MacroMeta models shown in chapter 2 of this paper.

The documentation situation is also slowly improving. The quality declarations, based on the European Quality Concept have become a relative success. Since a few years back they cover all official statistics in Sweden, not only the statistics for which Statistics Sweden is responsible. It is mandatory for all agencies in Sweden producing official statistics to submit an updated quality declaration to Statistics Sweden for all statistical products for which they

are responsible. During the first years, the quality of these quality declarations was varying – and the quality declarations produced by Statistics Sweden itself could not always serve as a pattern. However, the quality has clearly improved over the years and is now at least acceptable, and in some cases quite good. Internet exposure and clearly allocated responsibilities have encouraged responsible statisticians to do their best.

If thus the coverage of quality declarations is 100%, it is much lower for METADOK documentations (in the order of 50%) and complete SCBDOK documentations (in the order of 30%). In the case of METADOK, deficiencies in the software tools are one explanation. Among other things, the lack of intelligence in the existing METADOK tool necessitates quite extensive and time-consuming manual "proof-reading" and some undesirable duplication of work. Neither does the tool provide as many positive effects back to the producers, as would be desirable and possible. A new version of the tool is under development.

As for SCBDOK, one explanation for the low coverage so far is that the requirements on an SCBDOK documentation have been set relatively high: the quality of the documentation should be so good that for example a future researcher (living, say, 100 years from now) should be able to reuse and analyze microdata produced by the documented survey. A documentation of such quality could only be completed by a person with very good first-hand knowledge about the survey. Consequently, such documentation should not be postponed until after the survey has been completed; instead large parts of the documentation should be generated as a more or less automatic side-effect of the design process, as was discussed earlier in this paper.

5 LESSONS LEARNT, AND HOW TO MOVE ON

Experiences of metadata systems from Sweden and elsewhere have been summarized in a set of "golden rules"; Sundgren (2003a, 2003b). The rules are listed in figure 13 and consist of three groups aiming at designers, project managers/co-coordinators, and top managers, respectively. A good co-operation between these three categories of staff is absolutely essential for the success of a metadata undertaking in a statistical agency.

A proposal has been made to add the following rule:

- Metadata are as important as data, and metadata need as much work as data.

This rule underlines the importance of treating data and metadata management on an equal basis and in a well-integrated way. If we go back 50 years in time, when all statistics production was still done manually, without the help of computers, data and metadata management *were* actually integrated. In the paper questionnaires, metadata (the questions and instructions) and data (the filled-in answers to the questions) appeared together, and they continued to be treated together until the tabulation phase, where the data appeared as figures in tables, and the metadata showed up in the headings and labels. The first computers could only handle numbers and codes, not texts, and whereas the data processing became automated, the accompanying metadata were often "reinvented" by the operators of each production step: programmers, publication editors, etc.

If you are a designer…

- Make metadata-related work an integrated part of the business processes of the organization.
- Capture metadata at their natural sources, preferably as by-products of other processes.
- Never capture the same metadata twice.
- Avoid uncoordinated capturing of "similar" metadata – build value chains instead.[8]
- Whenever a new metadata need occurs, try to satisfy it by using and transforming existing metadata, possibly enriched by some additional, non-redundant metadata input.
- Transform data and accompanying metadata in synchronized, parallel processes, fully automated whenever possible.
- Do not forget that metadata have to be updated and maintained, and that old versions may often have to be preserved.

If you are the project co-ordinator…

- Make sure that there are clearly identified "customers" for all metadata processes, and that all metadata capturing will create value for stakeholders.
- Form coalitions around metadata projects.
- Make sure that top management is committed. Most metadata projects are dependent on constructive co-operation from all parts of the organization.
- Organize the metadata project in such a way that it brings about concrete and useful results at regular and frequent intervals.

If you are the top manager…

- Make sure that your organization has a metadata strategy, including a global architecture and an implementation plan, and check how proposed metadata projects fit into the strategy.
- Either commit yourself to a metadata project – or don't let it happen.
- If a metadata project should go wrong – cancel it; don't throw good money after bad money.
- When a metadata project fails, learn from the mistakes, and do it better next time.
- Make sure that your organization also learns from other statistical organizations.
- Make systematic use of metadata systems for capturing and organizing tacit knowledge of individual persons in order to make it available to the organization as a whole and to users.

Figure 13. Golden rules for the development and maintenance of metadata systems.

There are certainly metadata sets that need to be treated separately, as relatively autonomous resources of the data/metadata infrastructure of a statistical organisation; a corporate classification database is a typical example. However, in most situations data and metadata should be considered simultaneously. Thus when designing a process of a statistical system, one should consider where to get the necessary data and metadata from, and which data and metadata to produce as outputs from the process, and how these outputs should be taken care of.

Another important experience is summarised by yet another rule:

- View reality as it is – and not as it should be!

[8] Cf Porter (1985) and figure 1 earlier in this paper. Each process should add value to the data/metadata in an efficient way.

When you start a metadata project you always discover how ill co-ordinated the contents of existing statistical data are: how many different definitions there are for the same – or rather almost the same – concept, etc; some of the different definitions may be well motivated, but most of them usually are not; as is known from the theory of quality work, there is motivated and unmotivated variation. Metadata work and metadata tools have the good effect that they expose the lack of co-ordination very clearly and concretely. However, this should not lead to the conclusion that one should necessarily "tidy up the whole mess" before going on with metadata development. Instead one should at least temporarily accept the existing situation as it is and describe it as it is in the metadata. The metadata will then give us a very systematic and well-organised documentation of "the mess", and we will be much better off when starting to do something about it.

Another important insight from the development of data/metadata systems, especially when the output data and metadata are exposed on a website, is that a statistical agency have many important but also very different categories of users, with very different needs and different pre-knowledge about society in general and about statistical data about society in particular. Thus it is a demanding task for the statistics producer to provide a metadata infrastructure that could serve all these different users, without having to duplicate or even multiplicate the metadata infrastructure. Any kind of user of statistical data (researcher, analyst, journalist, politician, student, an ordinary citizen, etc) should be able to find out from the website of the statistical office, whether there are some statistical data available that could be relevant for him or her. It should be easy for the user to download possibly relevant data, and to interpret what they mean.

A data/metadata infrastructure serving a diversity of user needs, as just described, must not be based on any particular presumption about how users think, or how they would like to interact with the statistical system. The metadata infrastructure should be designed in a very general way, so that different applications and tools, serving different types of users, can easily be built on top of the infrastructure. These user-oriented tools and applications should communicate with the data/metadata infrastructure via standardised interfaces.

Thus an ideal statistical data/metadata infrastructure should be based on the principle of generality and diversity at the same time: general and standardised basic functions and metadata sets, supporting a diversity of user interfaces, user tools, and user applications via general, standardised interfaces.

No user should have to take courses in how to use the statistical system, before he or she can start using it. The user interface should have the same "look and feel" as other Internet-based systems, so that the user can use the system in an intuitive way. Whenever the user gets doubt about how to proceed, how to use a certain function, or how to interpret statistical data, different forms of help should present themselves automatically. Most users should not have to have any pre-knowledge about statistical terms, and neither should they have to cope with complex formal definitions.

On the other hand, the data/metadata infrastructure itself should be based on clearly defined concepts.

Future metadata systems should be as efficient as possible when it comes to filling the systems with metadata contents. Furthermore, efficient updating procedures have to be created, so that the metadata systems will be sustainable. In order to achieve these extremely important goals,

strong efforts have to be made in order to avoid “stand-alone”, manual processes in order to create and update metadata. Ideally almost all metadata should either be the result of automatic transformations of already existing metadata, or, if this is not possible, they should be by-products or side effects of other processes that have to be carried out anyhow. For example, definitions and descriptions naturally emanating from design processes should be automatically captured and organised by processes belonging to the metadata infrastructure of the statistical office. Similarly, the operation processes of statistical systems (data collection and data preparation processes, etc) should be designed so as to produce not only data, but also metadata, e.g. so-called process data informing about the performance of the processes themselves in terms of qualities and efficiency. Once again these more or less automatically generated metadata should be captured and organised by the metadata infrastructure. Metadata about errors and delays in the operation processes are examples of metadata that can be generated and systematically taken care of in this way.

Most importantly maybe, we should not have to wait five or ten years for the “ideal” data/metadata infrastructure to get ready. The system should emerge step by step. Intermediary results that are themselves useful for users and producers of statistical data should occur regularly. This does not happen by itself, but requires special attention in the planning and implementation of a data/metadata infrastructure.

On the other hand, one must not fall for the temptation to focus only on very visible and frequently demanded metadata for important end-users of statistics. If metadata systems for such purposes are built separately, without sufficient thinking in advance about how they fit into the infrastructure, and how they can be maintained with updated metadata in a sustainable and not too resource-consuming way, they may become a too heavy burden for the statistical office. Designers of statistical metadata systems must have a “split vision”: producing useful short-term results, without neglecting the long-term strategy.

6 STANDARDISATION IN INTERNATIONAL CO-OPERATION

Standardisation in international co-operation is an important part of our vision for the future of statistical data and metadata management. There are at least three relevant aspects of this co-operation:

- conceptual and contents-oriented standardisation
- technical standardisation
- methodological standardisation

In a modern society it is of utmost important that official statistics are comparable and coherent, internationally as well as nationally. International standardisation of concepts and contents is a key to achieving these objectives. Compliance with standards will never be perfect – there are many reasons, even some good reasons, why it may not always be possible, or even desirable, to comply with standards. Nevertheless, the existence of recognised statistical standards will make it possible to describe rather precisely, how a particular practice deviates from the relevant standard, and this makes it easier for a user of the statistical data to determine if and how these data can possibly be compared with other data, despite the deficiencies of the data. Obviously such descriptions of discrepancies between the “ideal”, standardised definitions of concepts and contents, on the one hand, and the definitions actually used, on the other, are important parts of the metadata for the statistical data concerned.

If conceptual and contents-oriented standardisation promotes comparability and coherence, technical standardisation will facilitate interoperability between statistical systems, internationally as well as nationally. Technical standardisation will make it easier to exchange data and metadata between statistical organisations in general, and between national and international statistical organisations in particular. Furthermore, it would obviously be of great advantage for users of statistical data and metadata, if, for example, all websites of national and international statistical agencies had the same "look and feel". Contents-oriented and technical standardisation in combination would ultimately make it possible for users to retrieve and combine statistical data from all over the world, without having to bother about, where the data are actually physically stored. The "push" techniques that prevail today could then be replaced by more flexible and less resource-consuming "pull" techniques, that is, the producing organisations would not have to physically send (push) their data and metadata to some common place; instead every user (including international organisations) could pull the data together, when they are needed, making it possible to view all statistical databases all over the world as one common, virtual database. This approach would also solve some important security and confidentiality problems. Today many statistical offices are prevented by law from making sensitive statistical data available to researchers in other countries – primarily because their own country's jurisdiction does not apply to these other countries, so even if contracts are written between the statistical agency and the researcher, the statistical agency would not be able to enforce any sanctions, should the researcher violate the contract.

The primary purpose of methodological standardisation is to promote quality and efficiency. It is well known by most statistical offices today that they have a lot to gain by activities such as benchmarking and development and implementation of best practices. Some statistical agencies have even been able to develop software together, and many more have found it useful to acquire and use software developed by other offices.

It is important that standards as regards statistical metadata are general in the sense that one and the same standard should cover all kinds of official statistics, that is, it should cover

- different topics (also called: subject matter areas, data categories, products, …)
- different structures of data (time series, cross-sectional, …)
- different aggregation levels (micro, macro, …)
- different regional levels (regional, national, international, ...)
- different stages in the production/dissemination chain (data collection, data preparation, aggregation and estimation, dissemination, analysis, …)

The metadata needs of international organisations are often special cases (specialisations) of the metadata needs of national statistical institutes, and any standardisation effort not starting from the needs of national statistical institutes will be harmfully limited in scope. It is much more difficult (and very seldom successful in practice) to start a standardisation process from special cases and generalise from those. There is much more hope for success and broad acceptance, if a standardisation process starts from a general case and makes sure, all the time during the process, that the general analyses and solutions are applicable to (and cover the needs of) all important special cases.

For example, one should not start by developing a metadata standard for some branch of economic statistics and then check if this standard could possibly be generalised to other branches of economic statistics or even to branches of social statistics. Such an approach is very time-consuming and expensive (in the long run; the first step will often seem to be simple

and inexpensive), and it is very seldom successful, mainly because it will finally become very complex and incoherent, if it will materialise at all.

Instead one should take a generalised approach from the beginning and use special cases for testing relevance and feasibility. The development and testing should go hand in hand all along the process, but generality should be in focus all the time.

But how general should a standard be? There is always a trade-off between the simplicity of a very general standard/theory with very few basic concepts (e.g. mathematics, set theory) and a less general standard/theory with more and semantically richer concepts, making it easier for a human being to relate it to his/her own practical reality.

For our purposes (metadata standards for statistical organisations), we may consider to focus on the following levels of generalisation/specialisation, starting from the most general one:

- all kinds of data/metadata
- all kinds of statistical data/metadata (including e.g. statistical data/metadata in pharmaceutical organisations)
- all kinds of statistical data/metadata relevant for official statistics
- statistical data/metadata of a certain topic/type/form/..., as exemplified above

If we should focus on one of these levels, it is in our opinion the third level that we should choose: all kinds of statistical data/metadata in connection with official statistics, covering all topics, types, forms, etc. More general standards, e.g. ISO standards, should be used wherever applicable. On the other hand it should be checked that solutions/standards proposed for official statistics in general could also be applied to special cases covering special topics, types, etc.

In our opinion it is very important that on-going international efforts concerning standardisation of statistical metadata, such as the SDMX initiative, take these viewpoints into consideration. Otherwise there is a high risk that proposed metadata standards would not be accepted, in the first place, that they will not work, in the second place, and that they will not work efficiently, in the third place.

Another important concern in on-going international co-operation on statistical metadata is the lack of standardisation of concepts and terms. The situation is still a little like “the Tower of Bable”. The communication about statistical data and metadata is often confused, even among professionals in the field, because of an abundance of terminologies. However, differences in terms and terminologies need not be disastrous, as long as the underlying concepts are well understood, and understood in the same way by different people. Unfortunately, concepts can only be discussed and agreed upon by using words, i.e. terms, so the best thing we can do is probably to agree upon a set of basic concepts by describing them by means of alternative terms, “permitted synonyms”. One could maybe even agree upon “recommended terms”, first-hand alternatives, but synonyms should still be permitted – as long as one is sure that they are really synonyms, alternative terms for the same concept.

Standardised concepts (with accompanying, structured lists of recommended terms and permitted synonyms) should hopefully emerge from on-going international co-operation in a not too distant future. It is also realistic to assume that statistical offices could learn from each other concerning best practices in designing standardised data/metadata infrastructure. This

process is already underway, but it could clearly be taken further. A standardised data/metadata architecture and standardised software solutions could finally emerge.

8 CONCLUSION

In our vision, metadata form an integral part of a universal statistical system, or rather a universal network of co-operating statistical systems, where a user is able to find, retrieve, interpret, combine, and analyse statistical data from all over the world, wherever they are physically stored. The user should be able to express his or her problems and information needs to the system in any form suitable to him or her, and the system should be able to associate the user's problem descriptions, however incomplete, with possibly relevant data, or at least provide suggestions to the user about how to proceed. Explanations should be given by means of illuminative examples rather than by means of formal definitions and instructions that require the user to learn new terms, unfamiliar to him or her, or take lessons about how the producer has organised the data. Thus the system should adapt to the user rather than the other way around. The system should be self-learning and self-adaptive. Such techniques are currently being developed in the field of artificial intelligence, for example in connection with the development of text mining techniques used in search engines and natural language interpretation; cf NEMIS (2002).

In our vision, there is also a less spectacular but equally important part of future metadata management, a part that is played behind the scenes, and which provides the necessary metadata to the more visible functions in an efficient way. As has been discussed in this paper the production of metadata has to be based on sources that generate metadata as by-products of other necessary functions, e.g. design and operation processes of the statistical systems. Metadata should be systematically captured, organised, stored, transformed, and made available to other processes, and all this should be done as automatically as possible; manual interventions should be avoided, but when they are needed, they should be supported by adequate software tools. Furthermore, the processes generating, storing, and transforming metadata, thus supporting, directly or indirectly, user-oriented processes, should themselves be metadata-driven. This means, among other things, that the software-driven processes will adapt themselves automatically to changes in data (and accompanying metadata) without interventions by programmers or other human operators.

In summary, it is our hope and belief that future statistical systems, based on integrated data/metadata management, and developed in international co-operation and using statistical standards and modern techniques, will raise official statistics to new heights, providing users all over the world with relevant, high-quality statistics in an efficient way, giving tax-payers good value for their money.

9 REFERENCES

ECB&Eurostat (2004) *CMFB Issues paper on the IT tools for the European monetary, financial and balance of payments statistics.* Joint paper by the ECB Directorate General Statistics and Eurostat for the 27th meeting of the Committee on Monetary, Financial and Balance of Payments Statistics, Luxembourg, 29 - 30 January 2004.

Eurostat (2003a) *Definition of Quality in Statistics.* Eurostat Working Group "Assessment of quality in statistics", Luxembourg, October 2003.

Eurostat (2003b) *Standard Quality Report.* Eurostat Working Group "Assessment of quality in statistics", Luxembourg, October 2003.

Langefors, B. (1966) *Theoretical Analysis of Information Systems.* Studentlitteratur, Lund, 1966.

NEMIS (2002) EU *Network of Excellence in Text Mining and its Applications in Statistics.* http://nemis.cti.gr/.

Neuchâtel Group (2002) *Neuchâtel Terminology. Classification database object types and their attributes. Version 2.0.* Neuchâtel, Switzerland, 2002.

Nordbotten, S. (1960) *The computers and the future design of statistics production.* In Norwegian: "Elektronmakinene og statistikkens framtidige utformning". Proceedings of the meeting of Nordic statisticians in Helsinki 1960.

Nordbotten, S. (1966) *A Statistical File System.* Statistisk Tidskrift (Statistical Review), No. 2, Stockholm 1966.

OECD (2003) *Towards a Quality Framework for OECD Statistics.* Notes by Michael Colledge, Statistics Directorate, OECD, October 2003.

Porter, Michael E. (1985) *Competitive Advantage: Creating and Sustaining Superior Performance".* The Free Press, New York, 1985.

Rosén, B. & Sundgren, B. (1991) *Documentation for reuse of microdata from the surveys carried out by Statistics Sweden.* Statistics Sweden 1991.

Statistics Sweden (2001a) *Quality definition and recommendations for quality declarations of official statistics.* Reports on Statistical Co-ordination for the Official Statistics of Sweden, MIS 2001:1.

Statistics Sweden (2001b) *Documenting statistical surveys, observation registers, and production systems. SCBDOK 3.0 – User handbook.* Available only in Swedish: "Att dokumentera statistiska undersökningar, observationsregister och statistikproduktionssystem. Användarhandbok för SCBDOK version 3.0." Prepared by Bo Sundgren. Statistics Sweden 2001.

Statistics Sweden (2003) *Metadata in the microdatabase, version 1.30.* Prepared by Matthias Abelin and Elisabet Andersson. Statistics Sweden 2003.

Sundgren, B. (1973) *An Infological Approach to Data Bases.* PhD thesis, University of Stockholm and Statistics Sweden, 1973. Also published as "Theory of Data Bases", Petrocelli/Charter, New York, 1975.

Sundgren, B. (2000) *The Swedish Statistical Metadata System.* Statistics Sweden 2000.

Sundgren, B. (2001) *Documention and Quality in Official Statistics.* Paper presented at the International Conference on Quality in Official Statistics (Q2001), Stockholm, May 14-15, 2001.

Sundgren (2003a) *Developing and Implementing Statistical Metainformation Systems.* Deliverable D6 from EU project "MetaNet" (IST-1999-29093), June 2003.

Sundgren (2003b) *Strategies for Development and Implementation of Statistical Metadata Systems.* Invited paper for the ISI session in Berlin, 2003.

Statistics Sweden (2004) *A description of the central metadatabase for macrodata. Version 2.00.* Statistics Sweden 2004.

Sundgren (2004a) *Metadata systems in statistical production processes - for which purposes are they needed, and how can they best be organised?* Paper for the Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS) in Geneva 9-11 February 2004)

Sundgren (2004b) *Documentation Templates and Metadata Models at Statistics Sweden.* Paper for the meeting of the Eurostat Metadata Working Group in Luxembourg 17-18 June 2004.

United Nations (1995) *Guidelines for the modeling of statistical data and metadata. United Nations Statistical Commission and Economic Commission for Europe.* Prepared by Bo Sundgren, reviewed at the work sessions on statistical metadata organised by the UN/ECE in 1992, 1993, and 1994. United Nations, Geneva, 1995.