Communicating in time and space –
How to overcome incompatible frames of reference of producers and users of archival data

Keynote paper at EDDI 2011, 5-6 December 2011, Gothenburg, Sweden
Bo Sundgren

Professor Bo Sundgren
Stockholm University
Department of Computer and Systems Sciences (DSV)
Affiliated with Dalarna University, Department of Informatics
Board member of Gapminder, www.gapminder.org
Chief Editor, International Journal of Public Information Systems (IJPIS), www.ijpis.net
bosund@dsv.su.se, bo.sundgren@gmail.com, bsu@du.se
https://sites.google.com/site/bosundgren/

The general purpose of documentation and metadata

People use different types of data representations for communicating information among themselves. Information, or knowledge, about reality is in the minds of people. With the possible exception of thought-reading, direct communication of information from mind to mind is not possible. Instead communication between people takes place via data representations, created and sent by one human being, and received and interpreted by others.

Data representations may also be used by one and the same person to help remember information. For example, we make written notes, voice recordings, or photographs and video films of things we want to remember.

Thus we may say that data may be used for communicating information in space (between people) and in time.

However, there is a major problem that needs to be tackled. When we interpret data, we use the experiences and knowledge that we already have in our minds, the reference knowledge, or the frame of reference. A person's reference knowledge will change over time: we experience and learn new things, and we forget things. Furthermore, different people obviously have different frames of reference, more or less different depending on their respective social and cultural environments, and their different life experiences.

To be able to communicate correctly and efficiently in time and space, we need to compensate for these differences, or incompatibilities, in frames of reference in order to increase chances that the receiver will interpret a communicated message in more or less the same way as was intended by the sender. Note that there is no direct way of actually verifying that communication is successful in this sense.

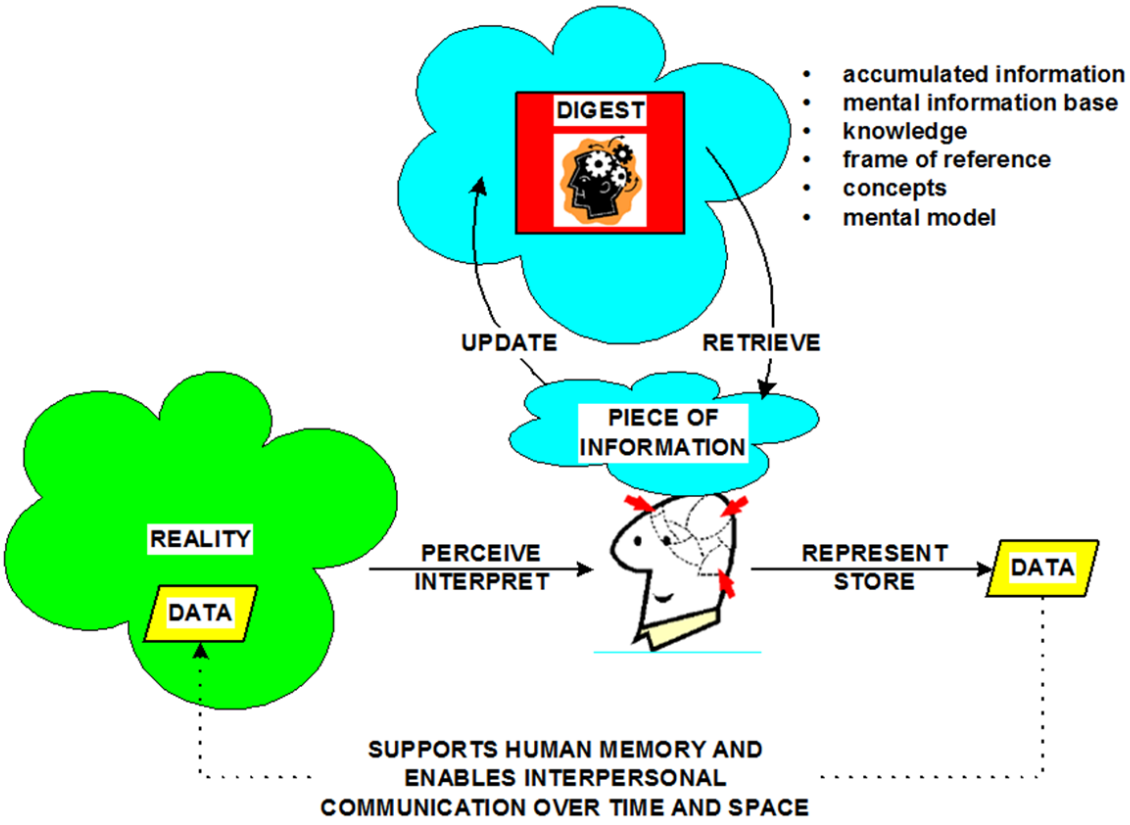

Figure 1. Reality – Information – Data.

In order to increase chances that receivers of data interpret the data in the way that was intended by the sender of the data, we may extend the data messages with metadata, data that describe and explain the meaning of the communicated data. Since the metadata are themselves data, they also need to be interpreted by the receivers, and these interpretations are of course also subject to errors and uncertainties. However, the metadata introduce some redundancy into the communicated messages, and hence hopefully decrease the variation and errors in the interpretations.

Documentation is a form of metadata. There is no sharp definition distinguishing between documentation and metadata. Generally speaking, documentation is typically more verbal and less structured than metadata, and documentation usually focuses on human interpreters, whereas metadata, especially structured and formalised metadata, are more adapted to computerised processing.

Metadata traditions

We may distinguish among at least the following major metadata traditions:

- The statistical tradition (from 1973 and onwards)
- The library tradition (e.g., Dublin Core)
- The archive tradition (DDI)
- Synthesis: business processes supported by information systems and a corporate data warehouse (see Figure 2)

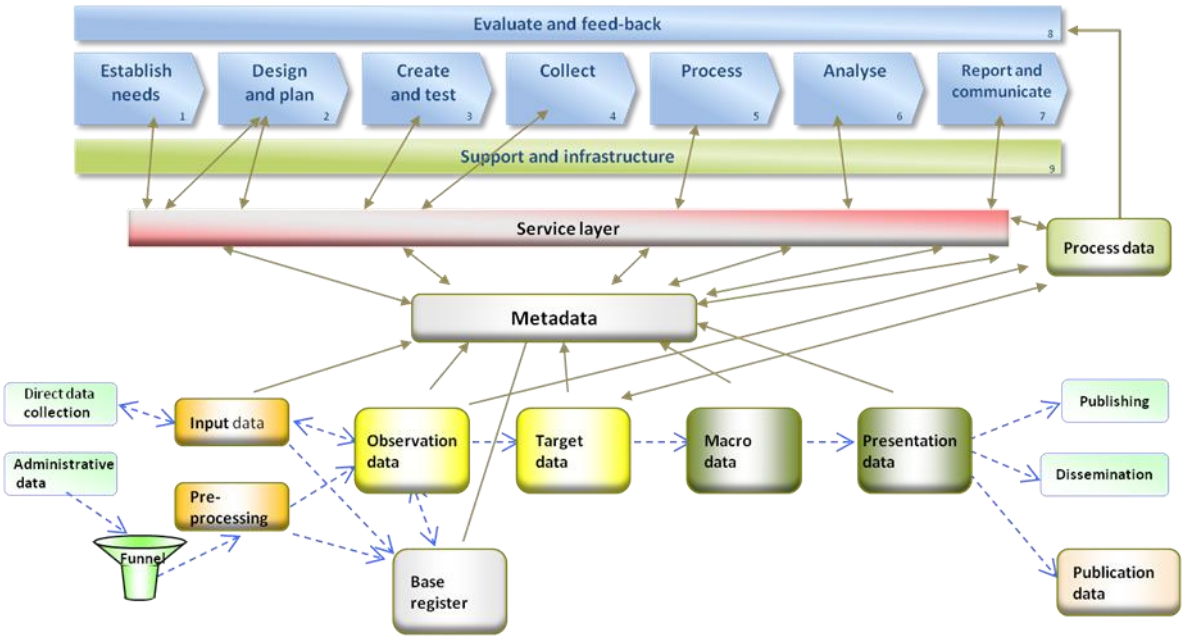

Figure 2. Conceptual view of the data warehouse of Statistics Sweden Lundell (2009).

The different metadata traditions now seem to be converging towards a model where metadata are generated, used, transformed, and reused in a natural way by processes in society and organisations, both business organisations and others, e.g., governmental agencies and research organisations. The processes are supported by a corporate data warehouse, where data and metadata are stored, transformed, and made available for different usages.

This model represents a drastic change from the classical archive tradition, where the archive was the final deposit for data that were to be preserved for future, rather infrequent, use and where data documentation was created for this archival storage, after the data had been used in a more active way.

The corporate data warehouse is a much more active form of deposit, where data and metadata are expected to play an important role as an integrated part of the processes of a society or an organisation, and where documentation and metadata processes are well integrated and synchronised with the business processes.

The statistical archive system

The corporate data warehouse model has a long history in official statistics. The concept of a statistical archive system, or a statistical file system, was created by Svein Nordbotten in the early 1960s, when he worked for the central statistical offices of Norway and Sweden. He later became the Director of the UN Statistical Office in New York, and still later professor of information science at the University of Bergen. He is still active as a researcher and advisor.

The archive-statistical principles can be summarised as follows:

- Reuse of existing raw data from administrative and statistical sources – for statistical purposes
- Continuous inflow of data (more or less)

- Data organised in a systematic way: statistical file system, databases, data warehouse
- *Ad hoc* production of statistics
- Systematic descriptions and definitions of data:
  - data and table definition languages; Nordbotten (1967): "*Automatic files in statistical systems*"
  - metadata; Sundgren (1973): *"An infological approach to data bases"*
- Standardised definitions and identifiers enabling flexible integration and combination of data: registers, classifications, standard variables
- Generalised software

See also:

The Ruggles Report (1965): "*Report of the Committee on the Preservation and Use of Economic Data*"

EU (2009): *"The production method of EU statistics -- a vision for the next decade"*

The history of the term "metadata"

The term "metadata" has created some controversy as regards its history. The most recent version of this history has been thoroughly researched and documented by Professor Jane Greenberg, Director of the Metadata Research Center, University of North Carolina, in her publication: Greenberg, J. (2010). Metadata and Digital Information. In *Encyclopedia of Library and Information Science, Third Edition*, 3610-3623. New York: Marcel Dekker, Inc. Her version of the history may be summarised as follows:

- The first known reference to "metadata" appears in Bo Sundgren (1973), *An infological approach to data nases*, pp. 104-105.
- Claims in the 1990s by Jack E. Myers to be the originator and owner of the term "metadata" were refuted by the U.S. legal system, with reference to Sundgren (1973) and "the longstanding use of the term in the statistical community."
- In 1986 Myers had registered "Metadata Inc." as a company, and "Metadata" as a trademark of that company. He later started to threaten people and agencies in the U.S. with legal actions, if they did not stop using the term "metadata" as a generic term.
- The Solicitor of the U.S. Department of the Interior decided that "Metadata" has entered the public domain by becoming a general term.
- Jack Myers has not been able to provide any documentation supporting his claim to have coined the term "metadata" in the 1960s.

The Data Documentation Initiative (DDI)

The roots of the Data Documentation Initiative (DDI) can be traced back at least to the 1980s. For a long time it was a typical representative of the archive tradition in the more orthodox sense, as described above, with the archive as a final deposit of data, and with documentation of the data as something created "afterwards," after the data had been created and used in a more active way, e.g., by a research process. The data documentation was strictly data oriented, typically organised in a so-called "codebook."

The more modern history of DDI can be found on the Web site of the DDI Alliance, www.ddialliance.org. During the last decade or so, the DDI model has been transformed into a life cycle model, which is much more in line with the archive-statistical model and the

corporate data warehouse model, as described above. More precisely, there are now two complementary documentation models maintained by DDI:

- DDI- Codebook (formerly DDI-2)
  - strictly data oriented
- DDI- Lifecycle (formerly DDI-3)
  - process and data oriented

DDI-Lifecycle is illustrated by Figures 3 and 4 from the Web site www.ddialliance.org.

Figure 3. What is DDI? – "A metadata specification for the social and behavioral sciences." – "Document your data across the life cycle."

Figure 4. What is DDI? – "Supporting the entire research data life cycle."

Life cycle models

Life cycle models are not new. Neither are they unique for information sciences. Here is a list of life cycle models from various contexts:

- Product life cycle (technical and marketing)
- Systems development life cycle (waterfall, etc.)
- Software development life cycle
- Business process life cycle
  - Generic Statistical Business Process Model (GSBPM)
- Data/metadata life cycle
  - Cycle de vie des données (CVD), Eurostat model
- Combined life cycle (combination of data and processes)
  - DDI-3: for the "social science business"

Life cycle architectures for statistical systems

Figure 5 illustrates an early version of a life cycle architecture for statistical systems. It was published in a Handbook by the United Nations and was based on earlier papers by Sundgren published from 1991 and onwards. Figure 6 shows a later version of essentially the same model – but with more modern process symbols.

In Figure 7 and Figure 8 the data warehouse has been introduced into the life cycle model, and control flows as well as data and metadata (and process data) flows and feedback loops are emphasized.

Figure 9 illustrates statistical systems as part of a gigantic societal feedback loop: *observations of society → statistical processing of observations → analysis → decision-making → implementation of decisions → changes in society → observation of effects*.

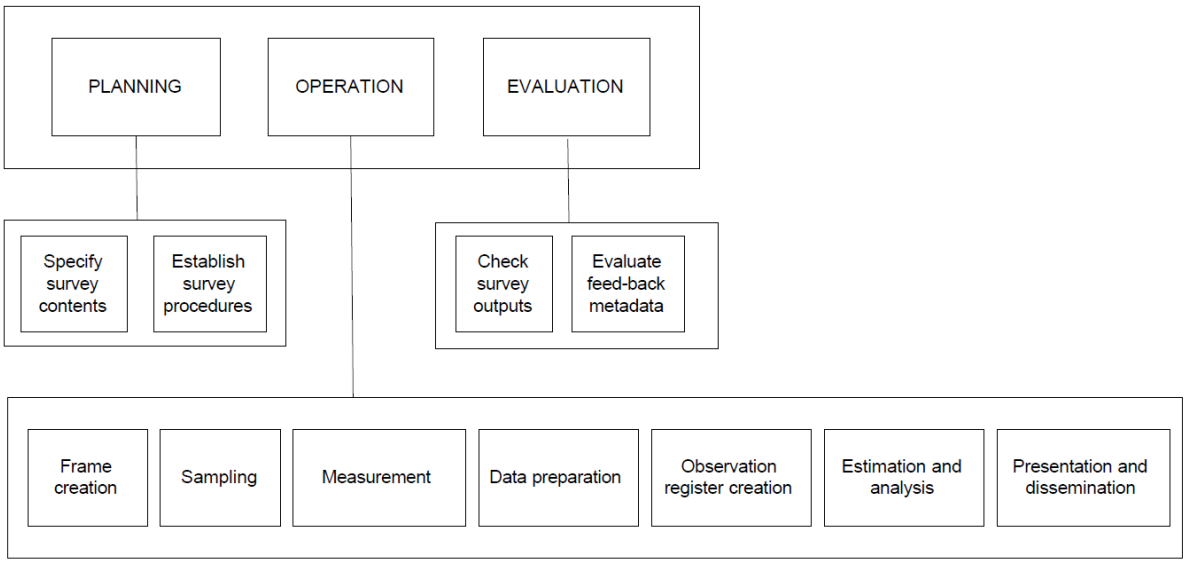

Figure 5. An early version of a life cycle architecture for statistical systems. Source: Sundgren (1999). "Information Systems Architecture for National and International Statistical Offices. Guidelines and Recommendations." UNECE, Geneva.

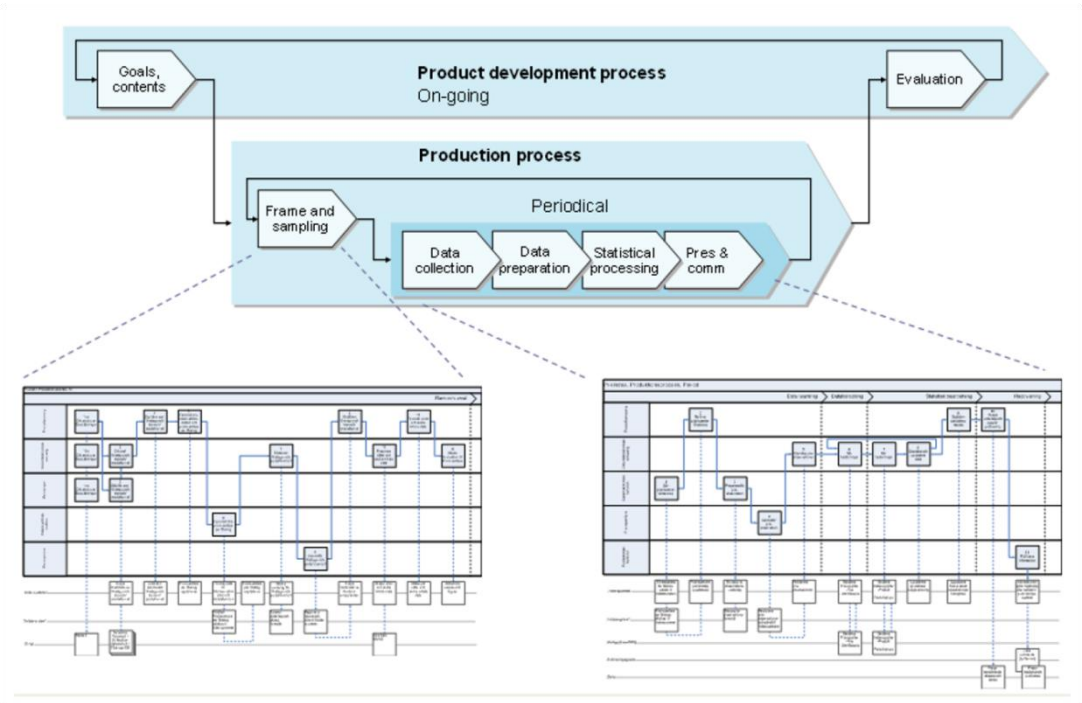

Figure 6. Statistics production: product development and production processes.
Source: Sundgren (2007). "Process reengineering at Statistics Sweden," MSIS Geneva.

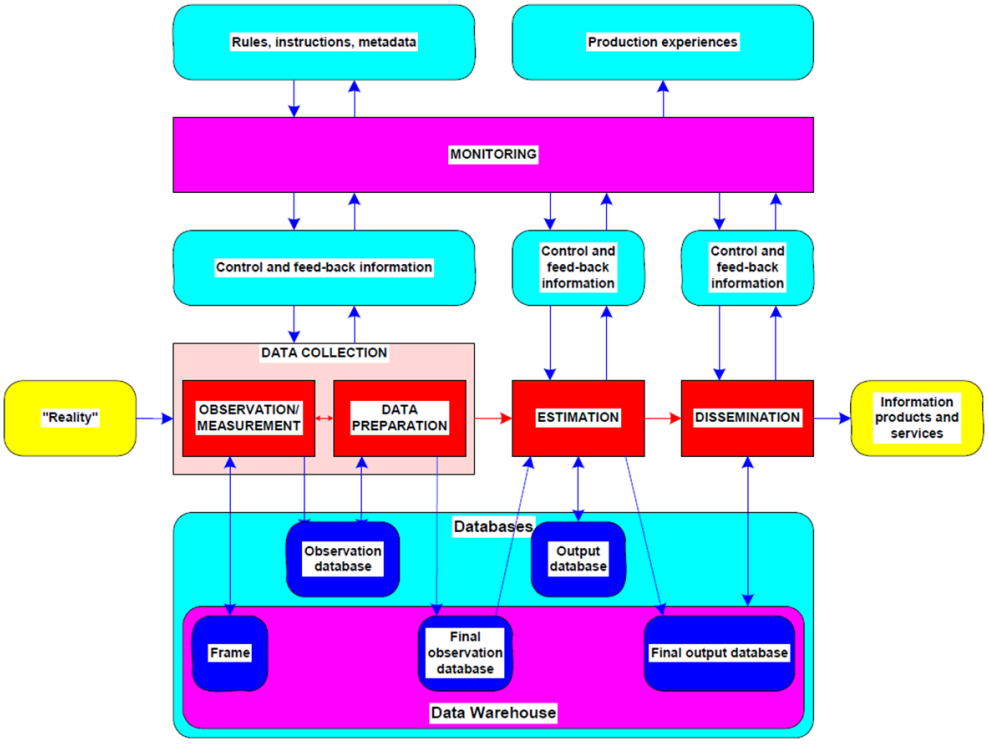

Figure 7. Basic operations in a database-oriented statistical system.
Source: Sundgren (2004b). "Statistical systems – some fundamentals."

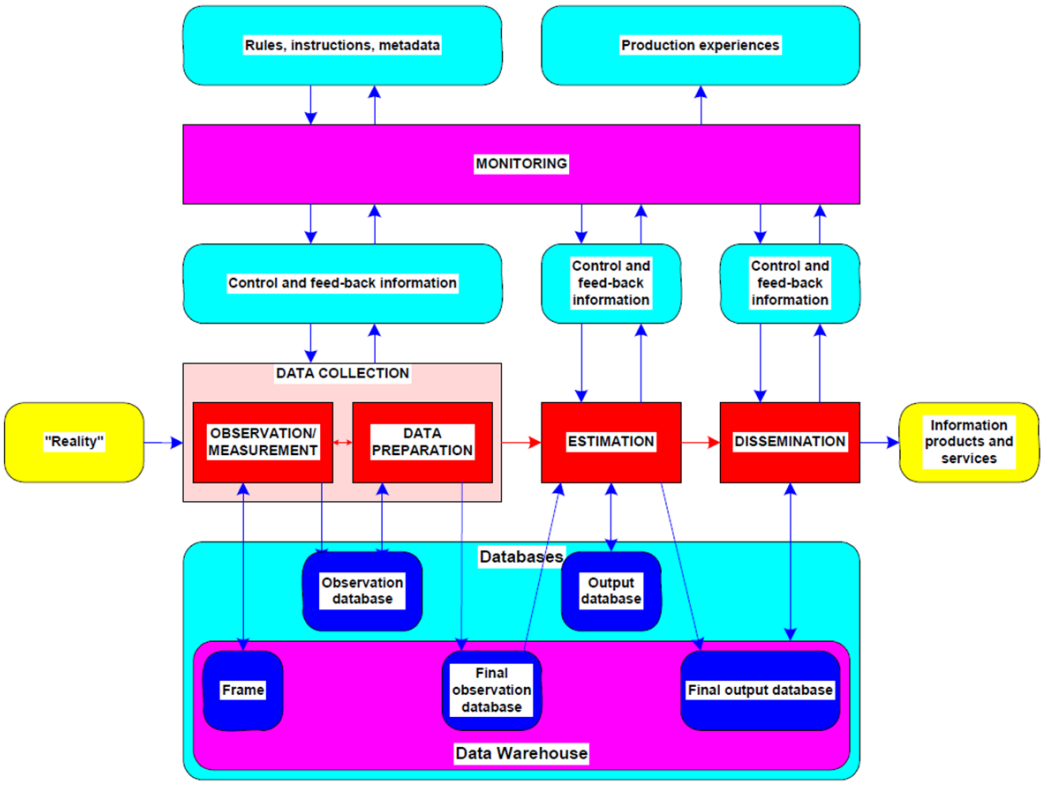

Figure 8. Control and execution of a statistical system.
Source: Sundgren (2004b). "Statistical systems – some fundamentals."

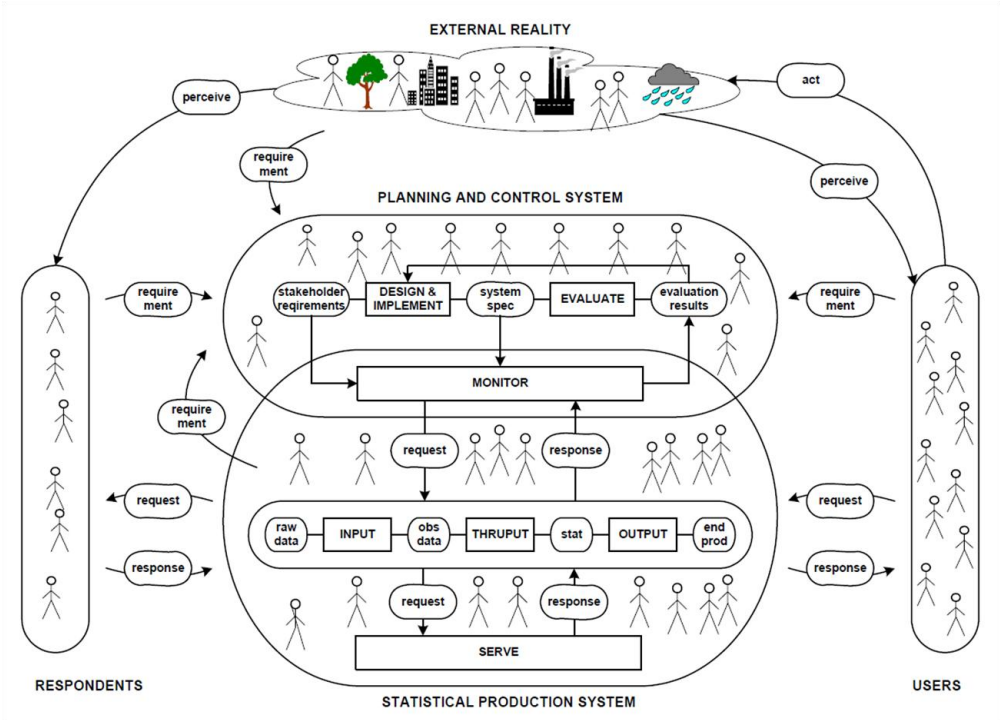

Figure 9. A statistical system in its environment.
Source: Sundgren (2004b). "Statistical systems – some fundamentals."

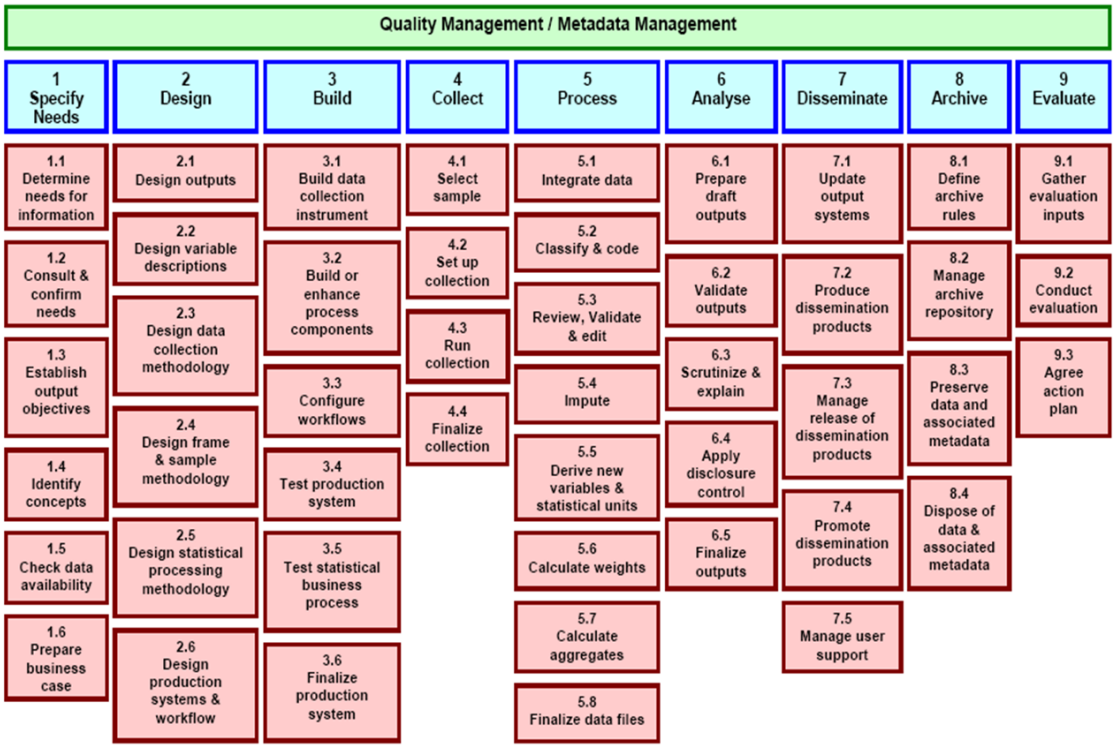

Figure 10. The Generic Statistical Business Process Model (GSBPM), levels 1 and 2. Source: UNECE (2009). "Generic Statistical Business Process Model." Geneva.

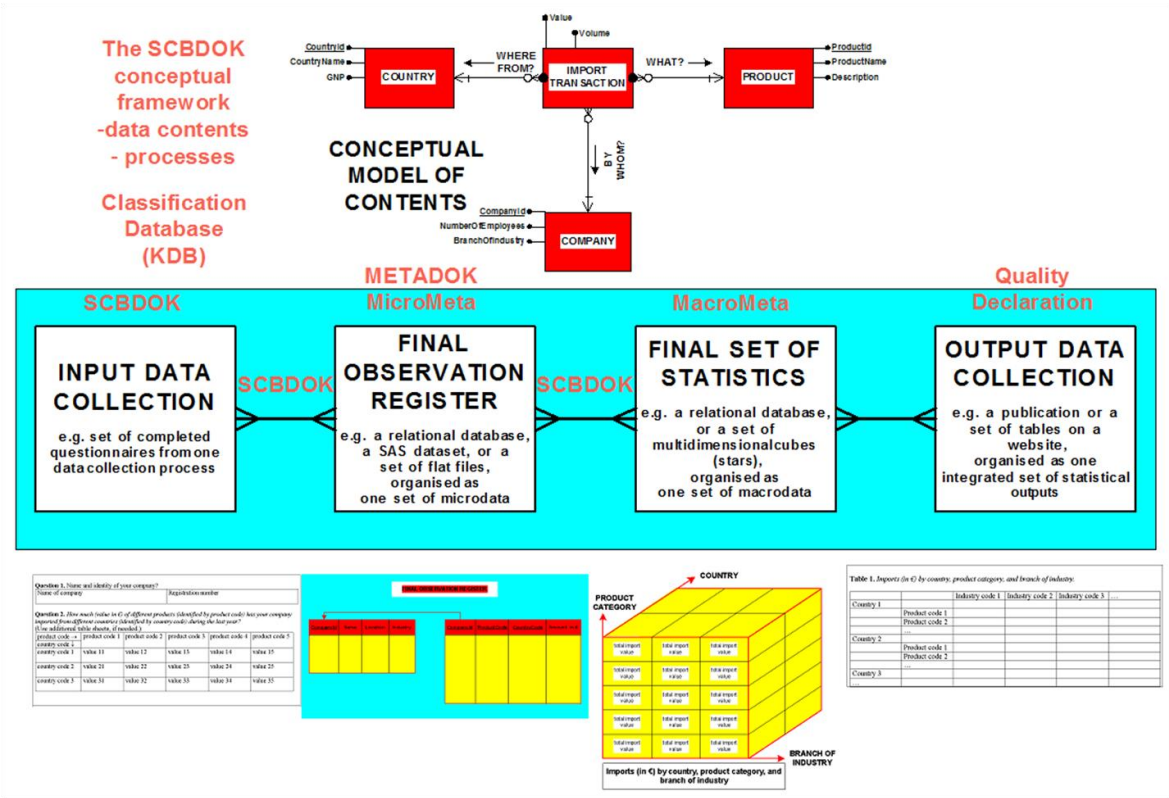

Figure 11. Focusing on data/metadata interfaces. Source: Sundgren & Lindblom (2004). "The metadata system at Statistics Sweden in an international perspective," Prague.

Figure 10 illustrates a contemporary version of the life cycle model of statistics production, the so-called Generic Statistical Business Process Model (GSBPM), originally developed by Statistics New Zealand in close cooperation with other statistical agencies in the world.

The life cycle model in Figure 11 focuses on the need for standard data/metadata interfaces between the major phases in statistics production, for example:

- Questionnaires (or equivalent data collection instruments)
- Final observation registers (microdata), possibly stored as relational databases
- Final sets of statistics (macrodata), possibly stored as multidimensional hypercubes
- Output data collections: statistical tables, graphs, etc.

Documentation templates

In order to capture and organise metadata and documentation emanating from the processes in statistics production, it is common to use structured documentation templates.

Figure 12 shows the SCBDOK documentation template, used by Statistics Sweden. It is a documentation template focusing on processes and final observation registers, aiming in particular at the needs of reusers of microdata.

Figure 13 shows the Quality Declaration Template as it was originally designed and used by Statistics Sweden. As can be seen, it focuses on the quality of the aggregated statistical outputs from statistics production, rather than on microdata and processes. This template, and the explanations behind it, also served as the pattern for Eurostat's first documents concerning the quality concept and quality reporting. See Eurostat (2003a) "*Definition of quality in statistics*," and Eurostat (2003b) "*Standard quality report*." Later updates are available. For detailed explanations of the Swedish Quality Declaration Template, and the concepts behind it, see

- Statistics Sweden (2001). "*Kvalitetsbegrepp och riktlinjer för kvalitetsdeklaration av officiell statistik - Quality definition and recommendations for quality declarations of official statistics*," MIS 2001:1, in Swedish, but contains a summary in English.

Data quality

The quality of statistical data is in the eye of the beholder, that is, the quality of statistical data is always dependent on what the data are going to be used for. The same data may be of good quality for one purpose and of bad quality for another purpose. Thus it is only the user who can finally determine if the quality of certain statistical data is good enough. However, the producer of statistical data can do a lot to make it easier for the user to judge the quality of data for a certain purpose. The producer may communicate a lot of valuable information for quality judgments through well elaborated process documentation and quality declarations.

Statisticians often find it easier to define quality in a negative way as "absence of errors," and it is true that it is easier to describe and measure errors than to give a direct description of data quality in a more positive way. The sources of errors occur all along the life cycle of the statistical design and production process. This is illustrated by Figure 14 and Figure 15.

SCBDOK 3.0			
0	**General information**	**1**	**Contents overview**
0.1	Subject matter area	1.1	Observation characteristics
0.2	Statistics area	1.2	Statistical target characteristics
0.3	Official statistics?	1.3	Outputs: microdata and statistics
0.4	Responsibility	1.4	Documentation and metadata
0.5	Producer		
0.6	Mandatory response?	**2**	**Data collection**
0.7	Secrecy		
0.8	Destruction rules	2.1	Frame and frame procedure
0.9	EU regulation	2.2	Sampling procedure (if applicable)
0.10	Purpose and history	2.3	Measurement instruments
0.11	Users and usage	2.4	Data collection procedure
0.12	General approach to implementation	2.5	Data preparation
0.13	Planned changes		
3	**Final observation registers**	**4**	**Statistical processing and presentation**
3.1	Production versions	4.1	Estimations: assumptions and formulas
3.2	Archive versions	4.2	Presentation and dissemination procedures
3.3	Experiences from the latest collection round		
5	**Data processing system**	**6**	**Logbook**

Figure 12. The SCBDOK documentation template.
Source: Sundgren (2001). "Documentation and quality in official statistics." International Conference on Quality in Official Statistics, Q2001.

Quality Declaration Template			
1	**Contents**	**2**	**Accuracy**
1.1	Statistical target characteristics	2.1	Overall accuracy
1.1.1	Objects[3] and population		
1.1.2	Variables	2.2	Sources of inaccuracy[4]
1.1.3	Statistical measures	2.2.1	Sampling
1.1.4	Study domains	2.2.2	Coverage
1.1.5	Reference time	2.2.3	Measurement
		2.2.4	Non-response
1.2	Comprehensiveness	2.2.5	Data processing[5]
		2.2.6	Model assumptions
		2.3	Presentation of accuracy measures
3	**Timeliness**	**4**	**Coherence especially comparability**
3.1	Frequency	4.1	Comparability over time
3.2	Production time	4.2	Comparability over space
3.3	Punctuality	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 13. The Quality Declaration Template of Statistics Sweden.
Source: Sundgren (2001). "Documentation and quality in official statistics." International Conference on Quality in Official Statistics, Q2001.

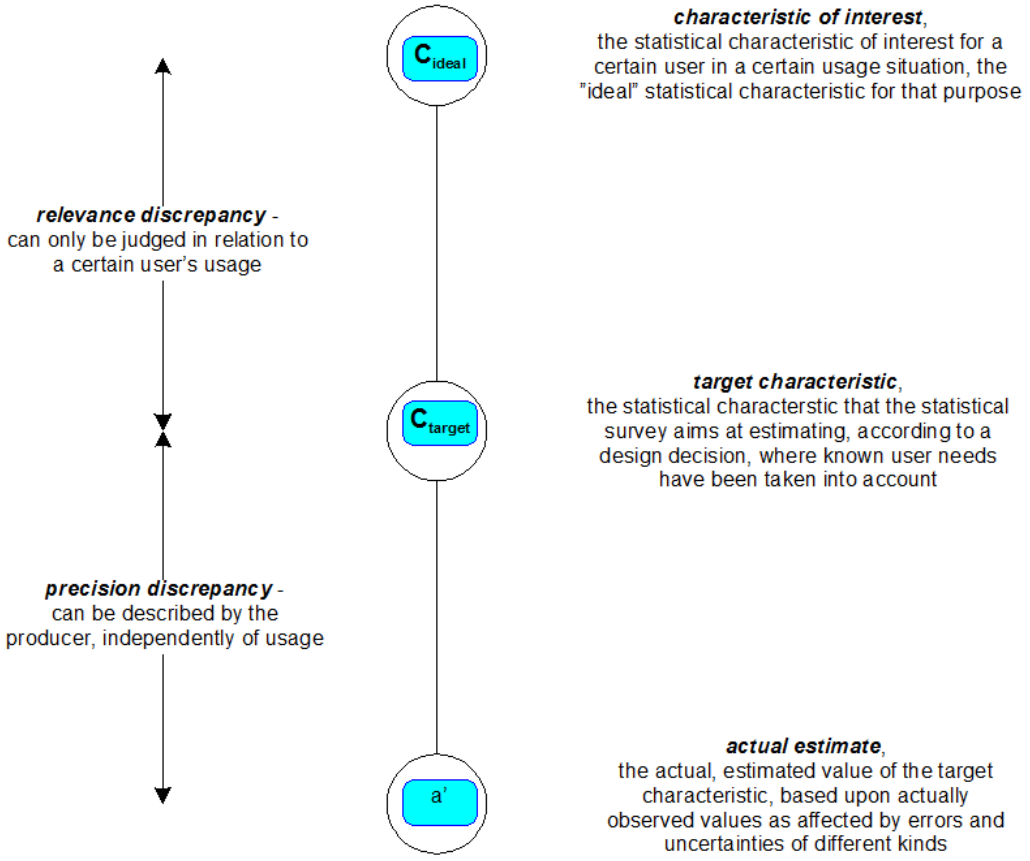

Figure 14. The quality of statistical data as affected by different discrepancies. Source: Sundgren (1995). Guidelines for the modeling of statistical data and metadata. United Nations Statistical Division, New York.

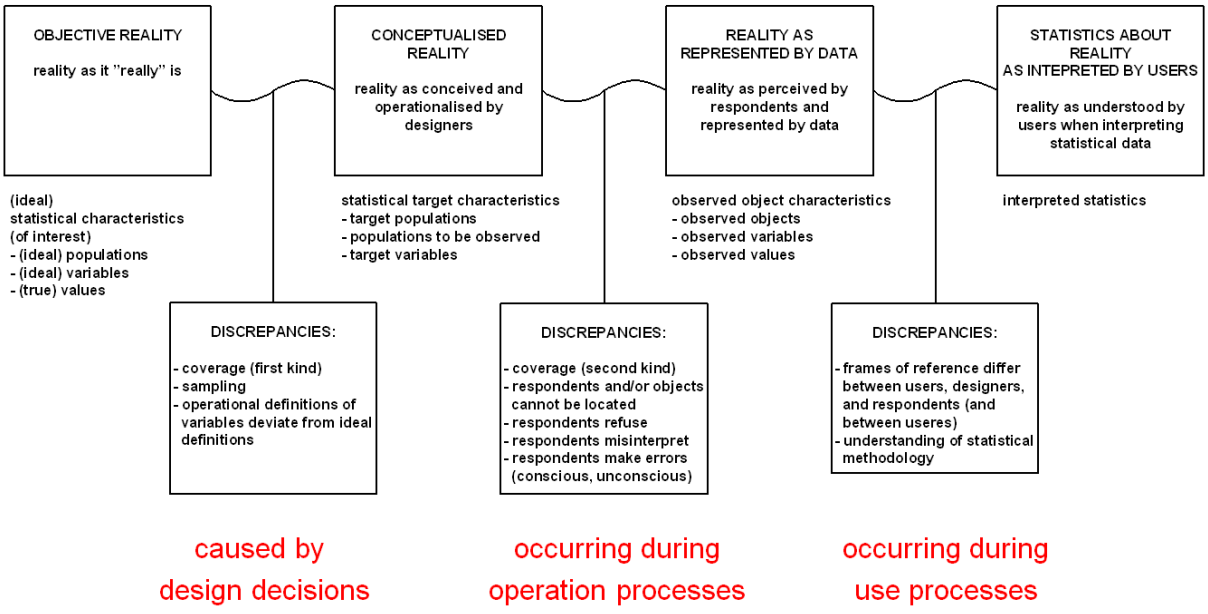

Figure 15. Discrepancies between reality "as it is" and as it is reflected by statistics. Source: Sundgren (2004b). "Statistical systems – some fundamentals."

As shown by Figure 14, the original wish of a certain user may be to obtain data measuring a certain ideal statistical characteristic, the **statistical characteristic of interest** for a certain purpose. However, there may be different users with different wishes as regards which statistical characteristic may be ideal for them. Moreover, it may be too costly, or not practically possible, to try to measure the ideal statistical characteristics desired by the users. During the statistical design process, discussions, trade-offs, and compromises will typically lead to decisions to aim for certain **target characteristics**, statistical characteristics that may not be ideal, but at least acceptable, for the users, and which are possible to measure in practice.

The target characteristic are not quite as relevant for the purposes that the users have in mind as the ideal statistical characteristics would have been. The discrepancy between an ideal characteristic and a target characteristic is called the **relevance discrepancy**, and it can only be judged in relation to a certain purpose.

However, not even the target characteristics, decided upon and defined during the initial design process, may be easy to measure. As is well known, and as is elaborated in more detail in Figure 15, there are a number of error sources, which will create another type of discrepancy, the so-called **precision discrepancy**, between the true values of the target characteristics and the estimated values of these characteristics, based on actually observed values, and as affected by different kinds of errors and uncertainties.

As illustrated by Figure 15, one may distinguish between the following types of discrepancies:

- Discrepancies caused by design decisions
- Discrepancies occurring during operation processes
- Discrepancies occurring during use processes

Figure 15 also gives some details about the nature and causes of these discrepancies.

"Statistical characteristics" vs. "statistics"

As should be obvious from the discussion that we have just had, it is essential to distinguish between "statistical characteristics" and "estimated statistical characteristics," or "statistics."

A **statistical characteristic** is defined as follows; see Sundgren (2004b):

- A statistical measure (m) applied on
- the (true) values of a variable (V); V may be a vector
- for the objects in a population (O)
  where
- O.V.m = statistical characteristic
- O.V = object characteristic
- V.m = parameter

Examples of statistical characteristics:

- Number of persons living in Sweden at the end of 2001
- Average income of persons living in Sweden at the end of 2001

- Correlation between sex and income for persons living in Sweden at the end of 2001

An **estimated statistical characteristic**, or **statistic**, is defined as follows; see Sundgren (2004b):

- An estimator (e) applied on
- *observed* values of an *observed* variable (V');
- for a set of *observed* objects (O') *allegedly* belonging to a population (O)

Ideally the value of a statistic O'.V'.e should be "close to the true value of the statistical characteristic O.V.m that it aims at estimating."

Examples of statistics:

- The estimated number of persons living in Sweden at the end of 2001
- The estimated average income of persons living in Sweden at the end of 2001
- The estimated correlation between sex and income for persons living in Sweden at the end of 2001

Conceptual models and object graphs

Conceptual models and object graphs are well known and powerful tools for defining and communicating the contents and structure of statistical data and metadata.

Figures 16-19 illustrate the following conceptual metadata models by means of object graphs:

- *Figure 16. The simplified MicroMeta model: Metadata overview for statistical microdata in final observation registers.*

- *Figure 17. The complete MicroMeta model.*

- *Figure 18. Simplified MacroMeta model. Metadata overview for multidimensional statistical macrodata in the online statistical database of Statistics Sweden.*

- *Figure 19. Complete MacroMeta model.*

Please use the zoom function in order to make the figures readable.

More details about these metadata models can be found in the following paper, and in references in this paper:

- *Sundgren & Lindblom (2004). "The metadata system at Statistics Sweden in an international perspective," Prague.*

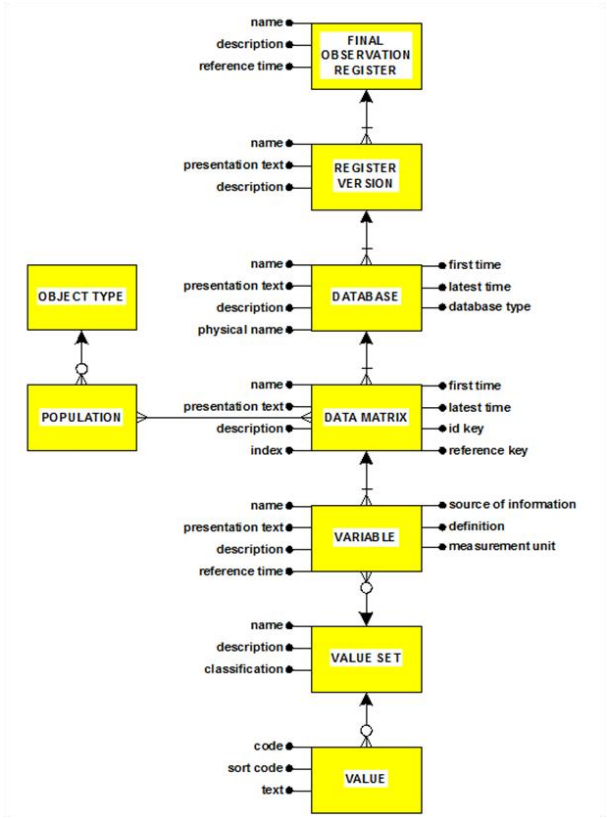

Figure 16. The simplified MicroMeta model: Metadata overview for statistical microdata in final observation registers. Source: Sundgren & Lindblom (2004). "The metadata system at Statistics Sweden in an international perspective," Prague.

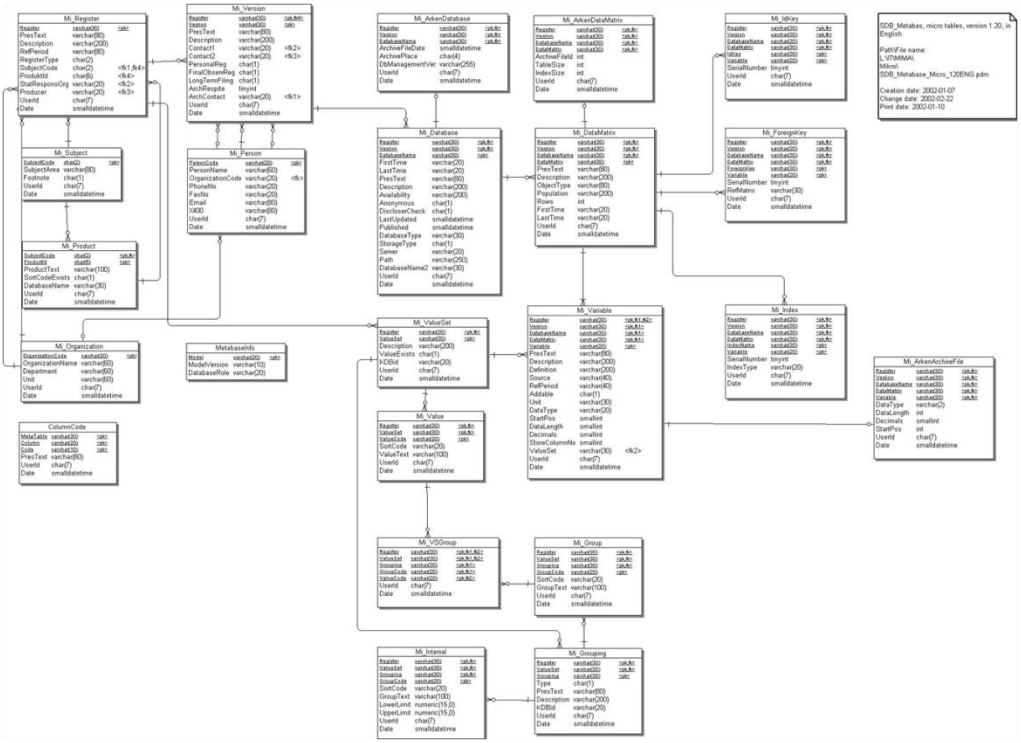

Figure 17. The complete MicroMeta model. Source: Sundgren & Lindblom (2004). "The metadata system at Statistics Sweden in an international perspective," Prague.

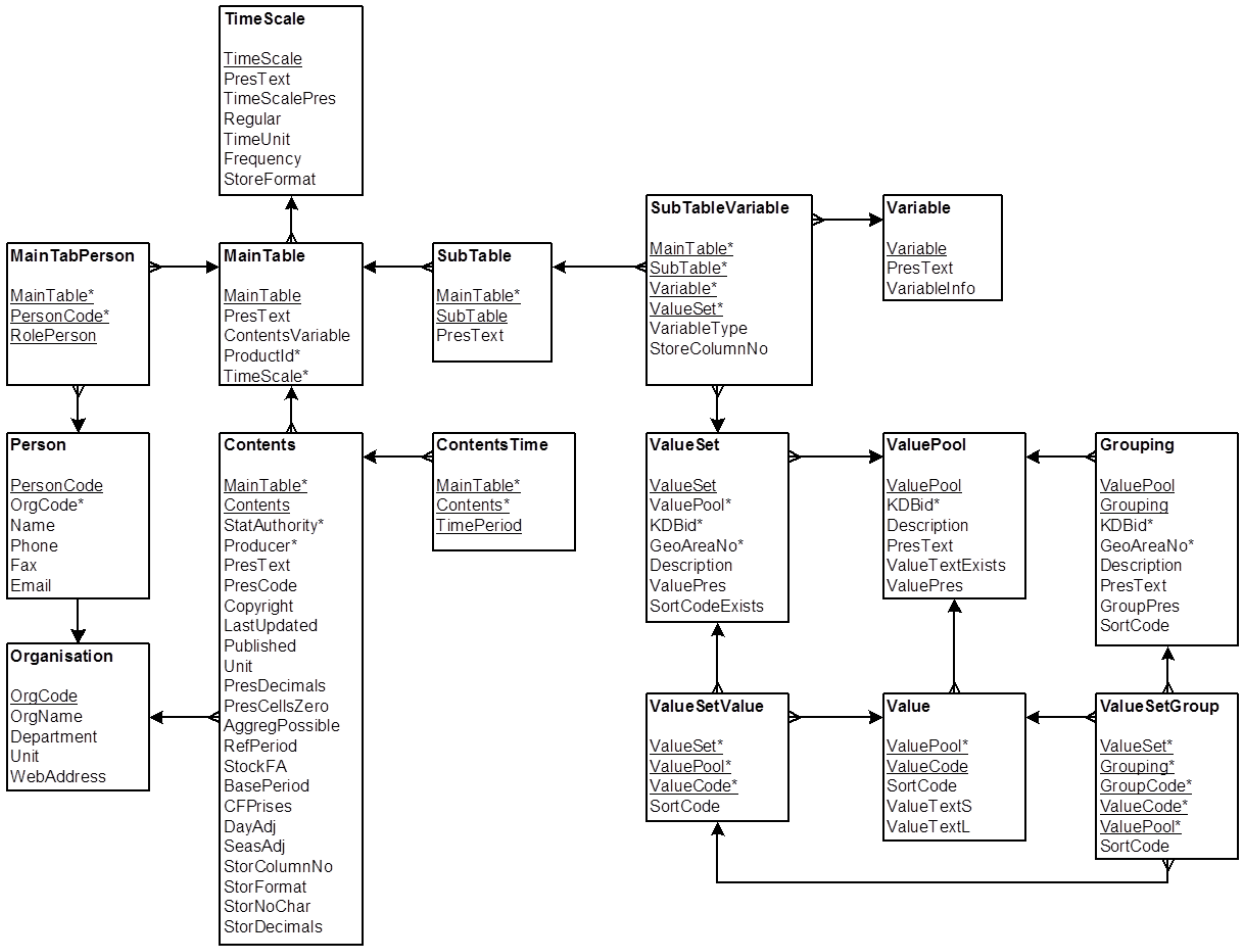

Figure 18. Simplified MacroMeta model. Source: Sundgren & Lindblom (2004). "The metadata system at Statistics Sweden in an international perspective," Prague.

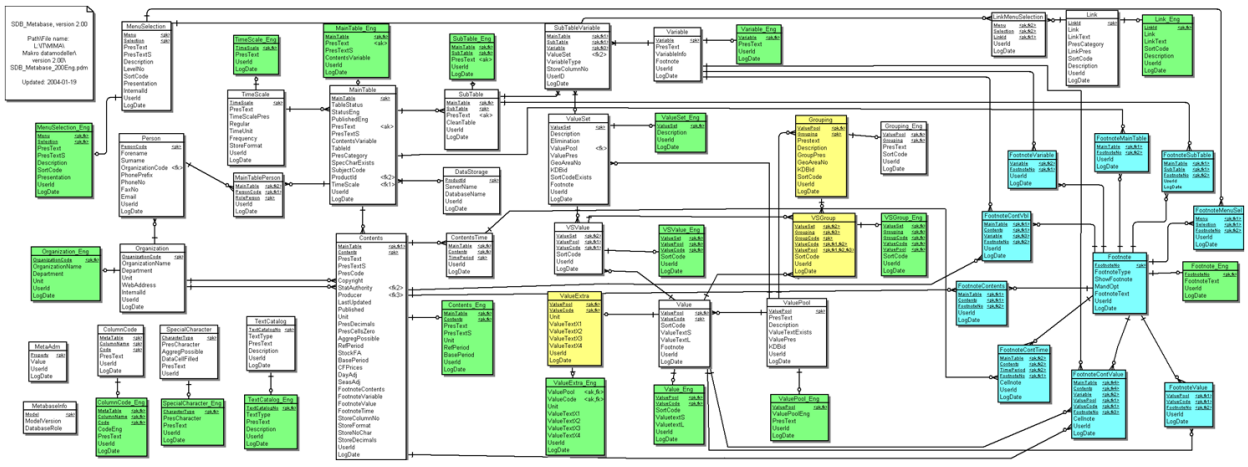

Figure 19. Complete MacroMeta model. Source: Sundgren & Lindblom (2004). "The metadata system at Statistics Sweden in an international perspective," Prague.

Success factors for documentation and metadata

There are three critical success factors for documentation and metadata in statistical systems:

- **Motivation**: How to respond to the arguments against documentation-related work?
- **Contents**: Which metadata are needed by which stakeholders for which purposes?
- **Management**: How to manage a metadata system in an efficient and sustainable way?

Arguments against documentation

Some common arguments against documentation-related work are:

- Time and costs: "We don't have the resources"
- "We have more important things to do"
- Dull, not fun, not rewarding
- Competent key persons are scarce resources
- "I know everything – come to me and ask"
- "I don't want to lose my knowledge monopoly"
- "The users don't ask for documentation"
- Easy-to-use tools are not available
- Documentation is produced as a separate activity
- There are too many types of documentation, contents overlapping and not well motivated, duplication of work

First of all, it is a myth that documentation-related work is costly and time-consuming. An experiment at Statistics Sweden showed that at most 2% of the budget of Statistics Sweden would be needed to produce high-quality documentation for all statistical products produced by the agency. This estimate was made for first-time documentation made by people who were competent statisticians, but not familiar from the beginning with the products they were going to document.

In fact, most of the arguments are about motivation. It is a challenge to prove to everyone needed for the job that documentation-related work need not necessarily be dull and unrewarding. As a matter of fact even many of those who believe that they know a statistical product quite well will learn a lot, and they will find it interesting and fun to learn these lessons. They will also discover that many users would appreciate getting more documentation that is more informative with regard to their needs and better and more innovatively presented.

It is also important to organise the documentation work in an intelligent and efficient way, avoiding duplication of work, and avoiding an investment of effort and resources into documentation and metadata that are really not needed, or of minor importance.

Everyone must feel that "there is something in it for me."

Which metadata are really needed?

It is important to focus on metadata which are really needed by someone for good purposes. Figure 20 provides a structure for identifying metadata needs, starting from the different categories of stakeholders in official statistics – and in metadata about official statistics. Note

that there are many other categories of stakeholders than the actual users/customers of official statistics, and that there are many different categories of users/customers.

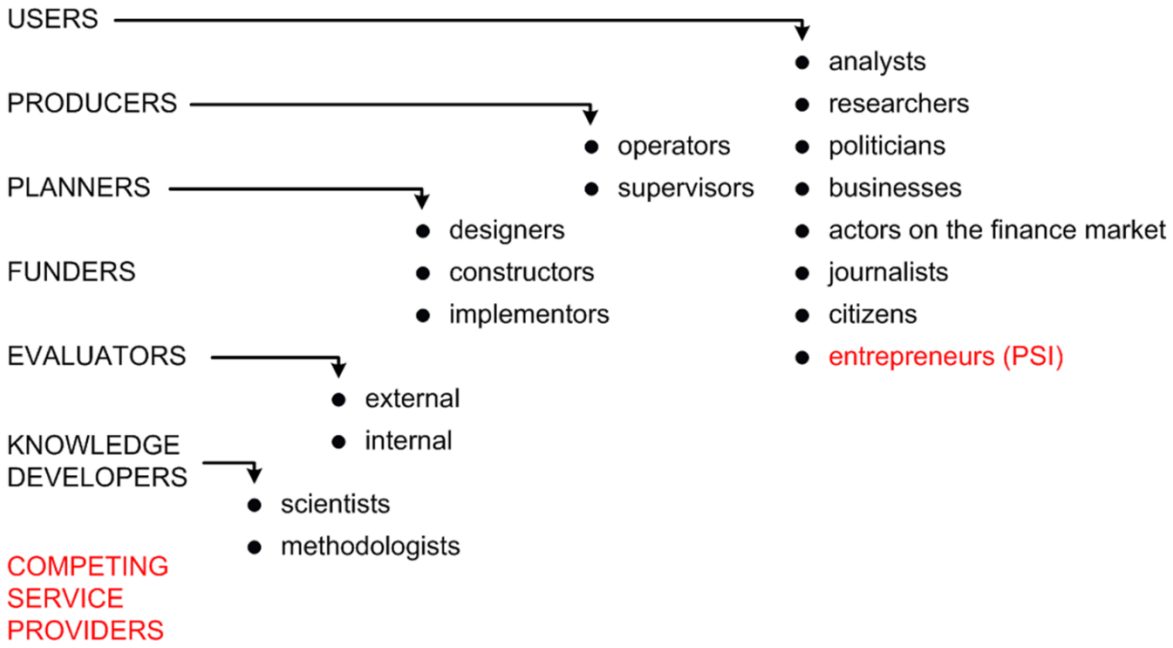

Figure 20. Stakeholders in official statistics – and in metadata about official statistics.

The next step is to identify the structure and contents of the metadata needed by the different categories of stakeholders. We will structure the needs by documentation/metadata objects and documentation/metadata variables (kinds of information informed about by the metadata).

Documentation/metadata objects

- **Datasets**
  - Information contents
    - Object types and populations
    - Object relations and relational objects
    - Variables and value sets
  - Physical datasets
- **Processes and systems**
- **Instruments and tools**
  - Methods, algorithms, programs
  - Questionnaires and other measurement instruments
  - Registers, classifications, other auxiliary datasets
  - Metadata, documentation

Documentation/metadata variables

- **For datasets**:
  - Definitions, verbal or formal
  - Quality variables, by quality component
  - Technical metadata, e.g., storage format
- **For processes and systems**:
  - References to input and output datasets

  - References to instruments and tools
  - Process data (paradata) generated by the processes
- **For instruments and tools**:
  - Documentation of instruments and tools
  - The instruments and tools themselves (*in extenso*)
  - References to systems and processes using them
  - User experiences

Managing a metadata system in an efficient and sustainable way

Figure 21 indicates important categories of documentation and metadata inputs and outputs for each one of the four major phases of the statistics production life cycle: (re)planning, execution, use, and evaluation.

The figure shows quite clearly where a certain type of documentation/metadata is produced, and where it can hence be captured into a well organised and easily accessible documentation/ metadata system. The figure also shows quite clearly how documentation/metadata, produced by one process, can be recycled and reused by other processes, possibly after some kind of transformation.

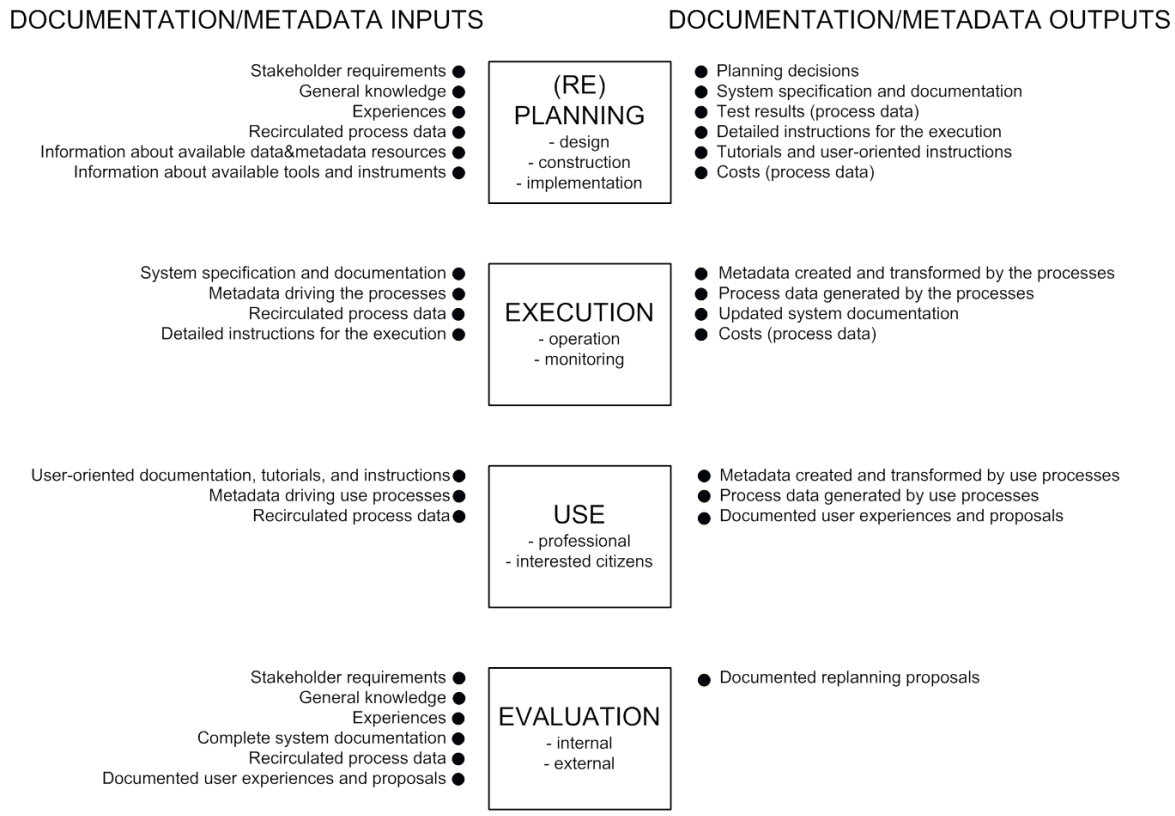

Figure 21. Use, production, and recycling of documentation of documentation and metadata.

The Swedish Statistics Commission

The Swedish government has set up a commission to investigate Statistics Sweden and the Swedish statistical system, and to work out proposals for improvements. The principal

investigator is Bengt Westerberg, a former minister of social affairs and leader of the Swedish Liberal Party, and he is assisted by 10 experts and a secretariat. The final report from the Commission is to be delivered by December 2012.

The Swedish government has given some 20 pages of directives to the Statistics Commission, containing a large number of important issues. The list of issues includes:

- Analyse centralised vs. decentralised system
- Examine in particular the quality and accessibility of statistics, including documentation, pricing, and confidentiality
- Analyse what it means for statistics production that state authorities as a rule should not sell goods and services on the market
- Analyse the impact on SCB of the PSI Act
- Propose measures to ensure and improve quality, accessibility, and documentation, including a monitoring system
- Propose a strengthening of SCB's cooperation with universities and other agencies
- Propose how the system of official statistics should be designed
- Propose changes in Swedish regulations as the result of new regulations and expected changes on the European level, e.g., the Code of Practice for European Statistics, the PSI Directive, and "the EU vision for official statistics"
- Propose any constitutional amendments deemed necessary

It can be seen from this list that issues concerning quality, accessibility, and documentation are felt to be major concerns for the future of Statistics Sweden and the Swedish statistical system. The Commission is also asked to propose changes in Swedish regulations that may be necessary because of new regulations and expected changes on the European level.

As one of the experts of the Statistics Commission, I have developed a number of proposals concerning quality, accessibility, and documentation. The proposals are grouped into four main categories:

- Independent monitoring system

- Free access to methods and tools (software, databases, registers, metadata, documentation, etc.) that have been developed for official statistics and funded by public money

- Continuously ongoing development of knowledge, methods, and general tools, funded by public money

- Systematic implementation of best practices

The EU vision for official statistics

Eurostat and the European Commission have outlined their vision for official statistics in the following document:

EU (2009). *Communication from the Commission to the European Parliament and the Council on the Production Method of EU Statistics: a vision for the next decade*. Brussels.

Here is a brief summary of the EU vision:

- **Current situation**: the augmented stovepipe model
  - Respondents asked for the same information more than once
  - Not adapted to collect data across domains
  - Little standardisation and coordination between areas
- **Demands for change**:
  - New information needs, often across domains, often ad hoc (e.g., in crises)
  - Decrease in reponse burden
  - Use of new ICT methods and tools to increase efficiency
- **Consequences on the level of Member States**:
  - Holistic approach, stovepipes replaced by integrated production systems around a data warehouse
  - Data obtained from existing administrative data and/or extracted directly from company accounts, combining survey data with administrative data, new efforts to ensure the quality of the data
- **Consequences on the EU level**:
  - Horizontal integration similar to the Member State level
  - Two elements of vertical integration: (i) collaborative networks, and (ii) direct production for the EU level, when there is no need for national data

It can be clearly seen that this vision is very much in line with the data warehouse-centred life cycle model that we have discussed in this paper, and which is also very much in line with ongoing developments as regards documentation and metadata management integrated with the basic processes of design, production, and use of statistical (and other) data. For example, systematic reuse of data and metadata, as well as integration across domains and stovepipes, are foreseen to become important features of the future European Statistical System.

Golden rules for metadata systems

As a summary of some main thoughts in this paper, I will present a set of "golden rules" for the design, development, and management of metadata systems.

The rules are formulated and elaborated in:

- Sundgren (2003a). "*Developing and implementing statistical metainformation systems*," Deliverable from EU project
- Sundgren (2003b). "*Strategies for development and implementation of statistical metadata systems*," ISI Berlin
- Sundgren & Lindblom (2004). "*The metadata system at Statistics Sweden in an international perspective*," Prague
- Sundgren (2004a). "*Metadata systems in statistical production processes – For which purposes are they needed, and how can they best be organised?*" UNECE/Eurostat/OECD, Geneva

I have derived these rules from experiences with metadata systems in Sweden and elsewhere. The rules are structured into three main groups of rules, aiming at designers, project managers/co-coordinators, and top managers, respectively:

Golden rules (1): If you are a designer…

- Make metadata-related work an integrated part of the business processes of the organisation.
- 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.
- 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 synchronised, 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.

Golden rules (2): If you are a project coordinator…

- 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 organisation.
- Organise the metadata project in such a way that it brings about concrete and useful results at regular and frequent intervals.

Golden rules (3): If you are a top manager…

- Make sure that your organisation 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. Lukewarm enthusiasm is the last thing a metadata project needs.
- If a metadata project should go wrong – cancel it; don't throw good money after bad money.
- When a metadata project fails, make a diagnosis, learn from the mistakes, and do it better next time.
- Make sure that your organisation also learns from failures and successes in other statistical organisations.
- Make systematic use of metadata systems for capturing and organising tacit knowledge of individual persons in order to make it available to the organisation as a whole and to external users of statistics.

References

EU (2009). *Communication from the Commission to the European Parliament and the Council on the Production Method of EU Statistics: a vision for the next decade*, Brussels.

Eurostat (2003a). *Definition of quality in statistics*.

Eurostat (2003b). *Standard quality report*.

Eurostat (2005, 2007). *European Statistics Code of Practice for the National and Community Statistical Authorities.*

Greenberg, J. (2010). "Metadata and Digital Information." In *Encyclopedia of Library and Information Science, Third Edition*, 3610-3623. New York: Marcel Dekker, Inc.

Lundell, L-G (2009). *Strukturerade datalager för effektivare produktion*. In Swedish. Statistics Sweden.

Lundell, L-G (2009). *Data warehouse for efficient statistics production*. In English. Statistics Sweden.

Nordbotten (1967): *Automatic files in statistical systems*

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

Ruggles Report (1965): *Report of the Committee on the Preservation and Use of Economic Data*

Statistics New Zealand (2004). *End-to-end business model - Business model transformation Ssrategy*. Statistics New Zealand.

Statistics Sweden (2001). "*Kvalitetsbegrepp och riktlinjer för kvalitetsdeklaration av officiell statistik - Quality definition and recommendations for quality declarations of official statistics*," MIS 2001:1, in Swedish, but contains a summary in English.

Sundgren, B. (1973). *An infological approach to data bases*. Stockholm University and Statistics Sweden, Urval No 7.

Sundgren, B. (1993). *Guidelines on the design and implementation of statistical metainformation systems.* Report for the UN/ECE METIS Group, established within the programme of work of the Conference of European Statisticians.

Sundgren, B. (1995). *Guidelines for the modeling of statistical data and metadata*. Published as Guidelines from the United Nations Statistical Division, New York.

Sundgren, B. (1999a). *Information systems architecture for national and international statistical offices. Guidelines and recommendations*. Conference of European Statisticians Statistical Standards and Studies No. 51, United Nations.

Sundgren, B. (2001a). *The AlfaBetaGammaTau-model: A theory of multidimensional structures of statistics*. MetaNet, Voorburg, the Netherlands.

Sundgren, B. (2001d). *Documentation and quality in official statistics.* Paper presented at the International Conference on Quality in Official Statistics (Q2001).

Sundgren (2003a). *Developing and implementing statistical metainformation Systems*. Deliverable D6 from EU project "MetaNet" (IST-1999-29093).

Sundgren (2003b). *Strategies for development and implementation of statistical metadata systems*. Invited paper for the ISI session in Berlin.

Sundgren, B. & Lindblom, H. (2004). *The metadata system at Statistics Sweden in an international perspective*. Invited paper for the conference "Statistics – investment in the future," Prague, Czech Republic.

Sundgren, B. (2004a). *Metadata systems in statistical production processes - for which purposes are they needed, and how can they best be organised?* UNECE/Eurostat/OECD Joint Session on Statistical Metadata (METIS), Geneva.

Sundgren, B. (2004b). *Statistical systems – some fundamentals*. Statistics Sweden.

Sundgren, B. (2004c). *Designing and managing infrastructures in statistical organisations*. Statistics Sweden.

Sundgren, B. (2005a). *Modelling the contents of official statistics*. Meeting of the SDMX Group at the OECD in Paris 2005.

Sundgren, B. (2005b). *A conceptual model of society as reflected by official statistics*. Statistics Sweden.

Sundgren, B. (2005c). *Modelling statistical systems*. 55th Session of the International Statistical Institute, Sydney, Australia.

Sundgren, B., Androvitsaneas, C., & Thygesen, L. (2006). *Towards an SDMX User Guide: Exchange of statistical data and metadata between different systems, national and international*. Meeting of the OECD Expert Group on Statistical Data and Metadata Exchange, Geneva, 2006.

Sundgren, B. (2006). *Reality as a statistical construction – Helping users find statistics relevant for them*. Q2006, Cardiff, U.K.

Sundgren, B. (2007a). *Navigating in a space of statistical surveys of society*. ICES-III, Montreal. Short version.

Sundgren, B. (2007b). *Navigating in a space of statistical surveys of society*. ICES-III, Montreal. Long version.

Sundgren, B. (2007c). *Process reengineering at Statistics Sweden*. MSIS 2007, Geneva.

Sundgren, B. (2008). *Classifications of statistical metadata*. METIS 2008, Luxembourg.

Sundgren, B. (2010a). *A systems approach to official statistics*. Official Statistics in Honour of Daniel Thorburn, pp. 225–260.

Sundgren, B. (2010c). *Designing surveys and statistical systems - complex decision processes.* Paper submitted to the Scientific Council of Statistics Sweden. In English.

Sundgren, B. (2010d). *Statistical file systems and archive statistics*. Paper presented at the Nordic Statistical Meeting in Copenhagen, 2010.

Sundgren, B. (2010f). *Citizen-centric access to statistics*. Project proposal. Approved by Vinnova for the Nordic-Baltic research program on "Citizen-centric eGovernment services."

Sundgren, B. (2011a). *Towards a system of official statistics based on a coherent combination of data sources*. Paper presented at the ESRA conference in Lausanne.

UNECE (2009). *Generic Statistical Business Process Model*. Geneva.