To document statistical surveys, observation registers and statistical production systems

Manual to SCBDOK, Version 3.0

Bo Sundgren

2001-11-01

SCBDOK 3.0			
0	**General information**	**1**	**Contents outline**
0.1	Policy area	1.1	Observation parameters
0.2	Domain of interest	1.2	Statistical target parameters
0.3	Part of Official Statistics of Sweden	1.3	Output: statistics and microdata
0.4	Responsible person	1.4	Documentation and metadata
0.5	Producer		
0.6	Mandatory duty to supply data to the survey	**2**	**Data collection**
0.7	Confidentiality and processing rules for personal data		
0.8	Appraisal and disposal rules	2.1	Frame and frame procedures
0.9	EU regulations	2.2	Sampling procedures
0.10	Objectives and background	2.3	Measurement instruments
0.11	Use of the statistics	2.4	Data collection procedures
0.12	Design and implementation	2.5	Data preparation
0.13	Planned modifications in future surveys		
3	**Final observation registers**	**4**	**Statistical processing and presentation**
3.1	Production versions	4.1	Estimation: assumptions and calculation formulas
3.2	Long-term (archive, terminal) storage versions	4.2	Presentation procedures
3.3	Experiences from the latest survey round		
5	**Data processing system**	**6**	**Log file**

CONTENTS

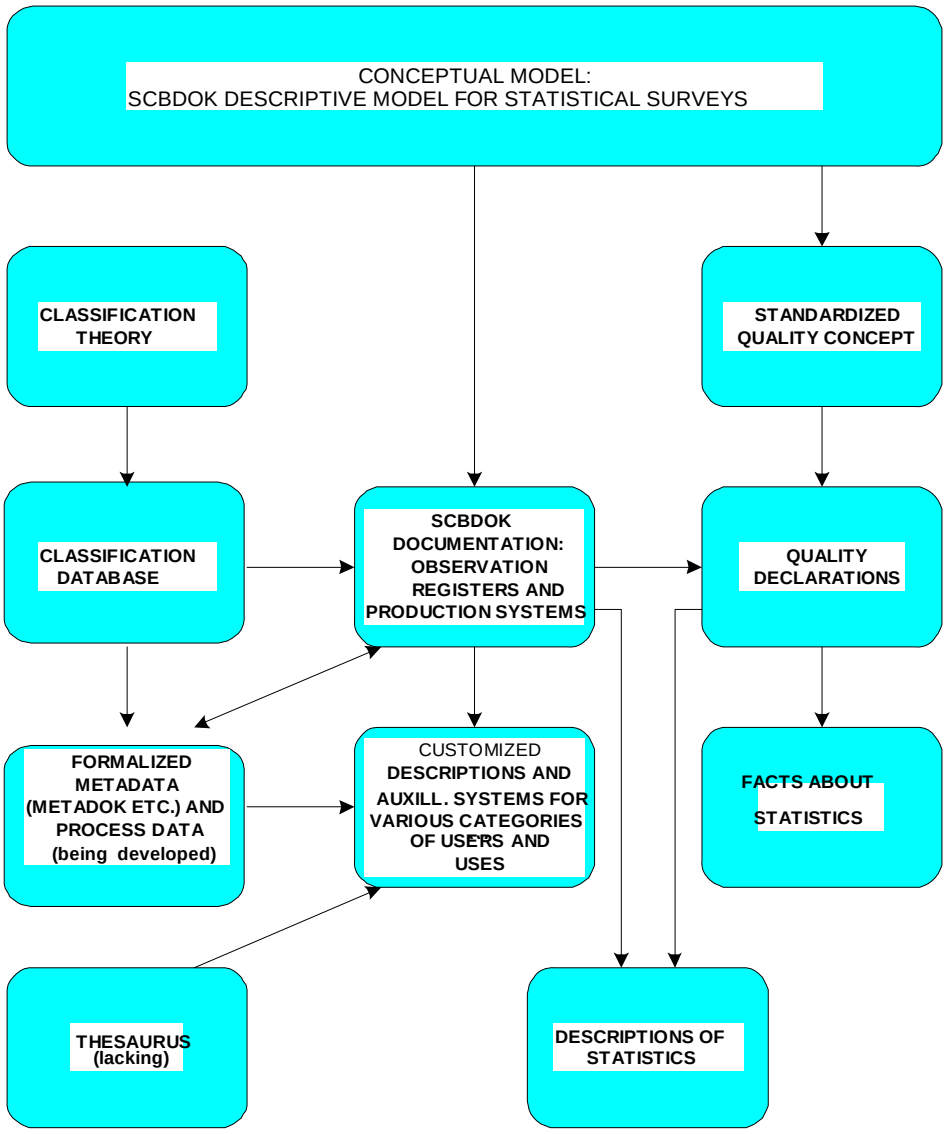
To document statistical surveys, observation registers and statistical production systems

Manual to SCBDOK, Version 3.0

Part 1: To document: Why? What? How?

Bo Sundgren

2001-11-01



Part 1: To document: Why? What? How?

1 Objectives: Why document?

There are three main categories of users and uses for an SCBDOK documentation:

1. *Users of statistics (macrodata) and statistical primary data (microdata).* The users need the documentation *both*
   - to identify relevant statistical data (microdata and macrodata) – assuming they have a subject-matter problem, and
   - to be able to interpret and use the statistical data they find.
2. *Producers of statistics*. The producers need the documentation to be able to use, maintain and develop their production systems, and to train new staff.
3. *Metadata-driven software* uses the documentation to produce, make accessible, analyze and present statistics and statistical primary data.

1.1 Documentation for statistics users

The main objective of SCBDOK is to facilitate the documentation of the final observation registers that originate from statistical surveys and that are stored for reuse in a more or less distant future. Researchers and expert planners form an important target group. If we imagine a researcher a hundred years from now who wants to reuse an observation register we have produced today by means of a statistical survey, we realize the high demands on such documentation. Some reasons are:

- The researcher/reuser cannot contact us who have participated in the implementation of the survey. We are all dead.
- The researcher will find it difficult to interpret the data included in the observation register. Which questions did the stored variable values provide answers to? Which was the target population? Did we try to observe all the objects/units in the population, or was it a sample survey? Which sampling strategy was employed, and which were the sampling probabilities for the various objects/units? Did we manage to observe all the objects/units in the sample or was there non-response? What did the non-response look like and how was it treated? Which other types of errors (e.g. measurement errors) occurred in the survey?
- The future researcher probably lives in a quite different society than we do. Even a good documentation might then be difficult to interpret, as definitions of concepts and discussions often rest on an implicit conception of society, how it works, which institutions there are, which administrative procedures that are employed etc. In some way we need to communicate the reference frame we take as self-evident to the future re-users of the data materials we leave behind.
- The technology to be used a hundred years from now may differ radically from that of today. The computers of today are perhaps found in museums, and most of our software has probably been long forgotten. Stored data will have to be reused with quite a different technology. How do we store and describe our data to make this possible at reasonable cost?

It is quite clear that a "myopic" documentation of the observation registers we leave behind is insufficient. A future re-user needs access to a detailed description of the procedures we have used to implement the statistical surveys that have resulted in the stored observation registers. Much of what we today consider self-evident must be described and explained to provide the future user with a correct reference frame. Any difficulties, problems, and deviations from plan that occurred during the implementation of the survey, might be very important to know in a re-use of the data.

Even users of final statistical tables might need a detailed documentation of the entire statistical production process to be able to assess the value of the statistics for various purposes and to analyze the statistics in a correct manner. For many users of aggregated statistics, the Quality Declaration (QD)[1] geared towards the final products is sufficient. The SCBDOK documentation and the QD are closely related. They are based on the same descriptive model, and a satisfactory SCBDOK documentation is easily adjusted and supplemented to yield a satisfactory QD. This also minimizes duplication of work.

Documentations and quality declarations are useful to the statistics users, not only in the use and re-use of statistical primary data and compiled statistics. They also have an important function in helping the user to identify which statistical data that are relevant to his/her issue (both microdata and macrodata). The documentations and quality declarations can then be used as a basis for free-text searches by search engines, preferably in combination with thesauri and other aids to efficient and "intelligent" searches.

The SCBDOK system and template were originally developed to document the final observation registers in such a way that their microdata should be re-usable after many hundred years, i.e. they were specifically intended to document observation registers that were to be archived. Such documentation doesn't necessarily include complete information on the production systems originally used. To a future user of microdata, most of the technical descriptions of today's production systems (Section 5 in the SCBDOK template) would be of little interest.[2]

In order to re-use microdata for other purposes than originally intended, it is strictly speaking not necessary to know the estimation procedures used in the original production.[3] But even if a re-user has a different purpose with his/her statistics production than Statistics Sweden originally had, it might of course be of great help to access the original methods and experiences in these respects. Consequently an SCBDOK documentation for re-use of microdata should include a description of the estimation procedures (Section 4 in the SCBDOK template).

However, an SCBDOK documentation might be regarded as a production documentation as well, i.e. as a documentation aiming to facilitate the work of the staff

[1] Quality Declarations are to accompany all Swedish official statistics. Cf. "Quality concepts and recommendations for quality declaration of official statistics", Meddelanden i samordningsfragor for Sveriges officiella statistik (MIS) 2001:1.

[2] Data storage constitutes an important exception. The long-term stored, final observation registers must employ simple, standardized and well-documented data formats that can be read and interpreted in the future as well. All technical information that is essential to reuse of stored data is to be found in Section 3 of the SCBDOK documentation.

[3] But it is important that the documentation covers all such metadata and process data that are needed if a reuser him/herself is to develop appropriate estimation procedures.

at Statistics Sweden responsible for planning, implementation and maintenance of the statistics production system. From this point of view it is evidently important that both the estimation procedures (SCBDOK Section 4) and the technical aspects (SCBDOK Section 5) are carefully described. Good production documentation is also of great use in training new staff.

It should be emphasized that the relatively detailed, process-oriented description of the data collection work specified in the SCDOK template, Section 2, is absolutely necessary even if we consider only the objective "re-use of microdata".[4] If we wish to provide a quality declaration of a final observation register that would permit it to be used in a more or less distant future, for more or less unknown purposes, it is indispensable to provide quite detailed information of how the data collection was carried out, which problems were encountered, how they were handled etc. Only then, a future re-user might hope to assess if and how the data can be used for his/her intended purpose.

A simple instance is how best to describe the definition of a variable. In variable lists and similar documents, we frequently try to provide short, abstract definitions of the variable meanings. But the best and most operational definition of a variable in a final observation register is given by the question originally put to the respondent, preferably together with the explanations and instructions, if any, that were available to the respondent and/or the enumerator.[5]

How, then, do we treat data from (external, administrative) registers? Well, such registers, too, are based on data collected from respondents, and in that connection questions have probably been put and answered with the help of various instructions and explanations. In such cases it might be necessary for the person doing the Statistics Sweden documentation to contact the institution responsible for the register and get information about the relevant administrative procedures and the original data source.

1.2 Documentation for statistics producers

To facilitate the use and in particular the long-term re-use of statistics is the most important objective of the SCBDOK documentation.[6] But there are other objectives. Even in the short term we quickly forget various aspects of the implemented surveys, facts that are important to know when we wish to interpret survey data, both the primary observations (microdata) and the final statistics (macrodata).

Today a survey is implemented by means of more or less complex technical and social systems, what we here term the statistics production system. Often we carry out new rounds of the same statistical survey month after month, quarter after quarter, year after year during a fairly long period. It is then rational to employ the same production system and successively develop it on the basis of achieved results and experiences. To do this as rationally as possible and to avoid becoming dependent on specific individuals, we need some kind of documentation of the production system. It must

[4] From the point of view of production documentation, the detailed descriptions of Section 2 are of course necessary.
[5] The simplest way to provide this type of definitions is to annex the questionnaire and instructions to the documentation, preferably in electronic form.
[6] Without adequate documentation, it rapidly becomes more or less impossible to reuse statistical data. Information capital is irreparably destroyed.

also include such technical information that is of minor importance when we only consider the facilities to interpret the published statistics and to re-use the final observation registers.

1.3 Documentation for metadata-driven software

Some parts of an SCBDOK documentation, primarily Sections 3.1 and 3.2, include very structured and formalized metadata of the type needed by various software to perform their functions. At Statistics Sweden, the METADOK system provides computer support for these parts of the SCBDOK documentation. By developing well-defined interfaces between METADOK and various software, it is possible to facilitate, coordinate and even make the metadata handling in statistics production fairly automatic.

When some software can perform its functions on the basis of existing metadata, stored in some metadata base, we talk about software driven by metadata. With such software, the user only has to indicate the data to be processed; he/she doesn't have to define them.

1.4 SCBDOK documentation as a fundament for other documentation

The SCBDOK documentation combines documentation for observation registers and documentation for production systems. Additionally it is a tool for those users of statistics who want a more thorough understanding than that provided by the Quality Declarations. To really understand published statistics, we have to know quite a lot about the processes behind the final figures. SCBDOK provides such information.

A carefully drafted SCBDOK documentation covers the main part of the facts about the statistics production necessary for other documentation purposes as well, such as the Quality Declarations and customized documentations for various categories of users and uses. In other words, the SCBDOK documentation provides a fundament for all other documentation and quality declarations. In turn, this is due to SCBDOK being based on a generic and detailed model of the statistical production process.

1.5 Summary of various documentation objectives

Objectives	Means
To identify statistics (macrodata) and microdata relevant to a subject-matter issue or domain of interest	Survey abstracts[7], thesaurus[8], classification database
To interpret statistics	Basic facts about the statistics[9], quality declarations, observation register documentations, customized documentations, and other means to facilitate the work of various categories of users and uses
To (re)use and combine microdata after short or long time periods	Documentations of observation registers, customized documentations, and other means to facilitate the work of various categories of users and uses; other users' analyses
Production and maintenance of statistics production systems, training of new staff	Production documentations
Customized processings and retrievals from the Statistics Sweden data storage	Formalized metadata for microdata and macrodata, metadata-driven software and production systems standardized metadata interfaces

Figure 1. *Objectives and means of documentation*

Figure 1 summarizes various documentation objectives and shows how these can be satisfied by various tools.

[7] By "survey abstract" is here meant the introductory part of what Statistics Sweden calls "Description of the statistics" (previously termed "Product description and quality declaration").
[8] Presently lacking.
[9] "Basic facts about the statistics" is a compulsory part in the template for Statistical Reports (SM). The target group is general users of statistics, i.e. non-experts. Experienced statistics users need complete Quality Declarations and the SCBDOK documentation

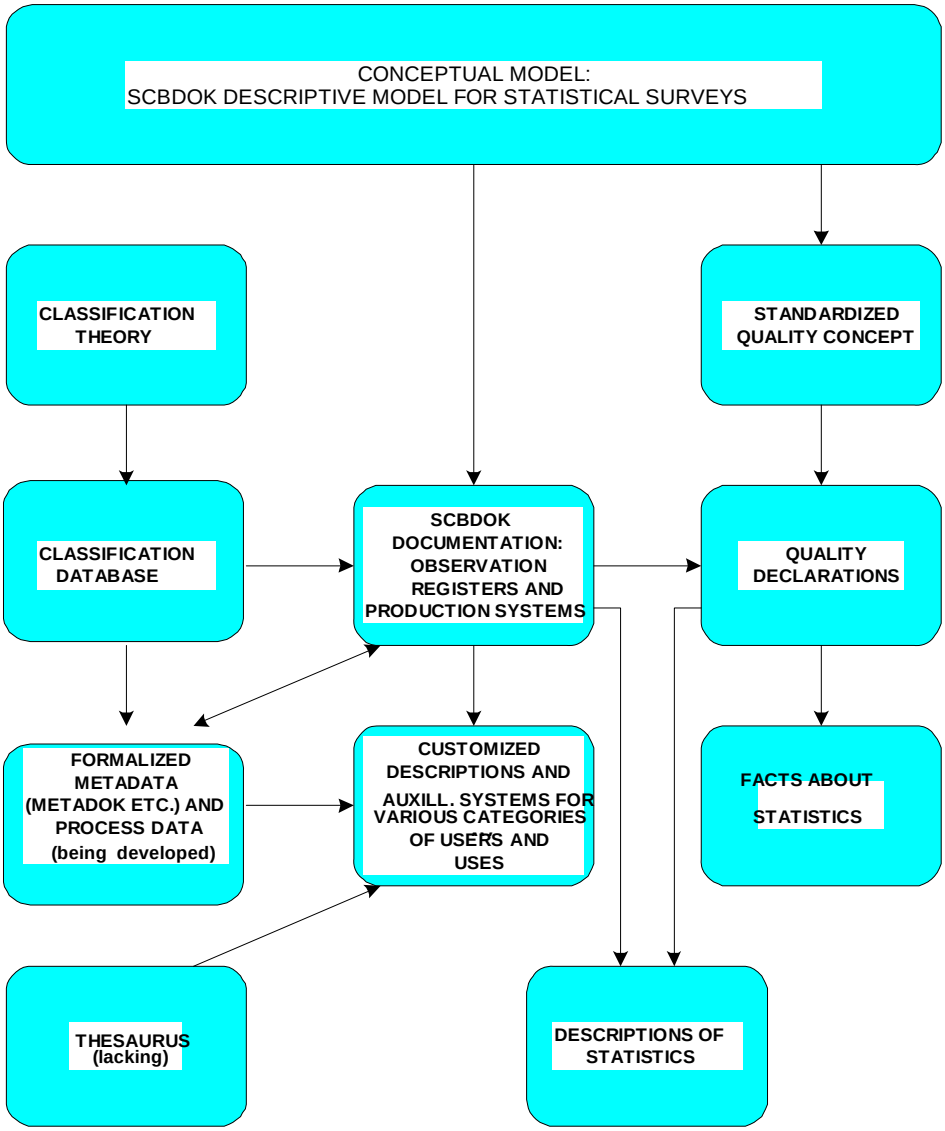


Figure 2. *Summary of different types of documentation and documentation tools and their inter-relationship at Statistics Sweden.*

2 Objects: What to document?

By “documentation object” is meant what the documentation “is about”. Some obvious objects for an SCBDOK documentation are

- observation registers, i.e. the data materials that are the result of a statistical survey (SCBDOK as observation register documentation)
- the statistical survey as a whole (SCBDOK as process documentation)
- the statistical production system (SCBDOK as production system documentation)

To achieve a complete documentation of these main objects, we must first document quite a number of other objects. Some examples are:[10]

Subject-matter/contents oriented objects

- observation parameters (including derived parameters)
- statistical parameters (parameters of interest, target parameters)
- object types, populations and domains of interest
- variables
- statistical measures (number, total, mean, percentage, correlation, etc.)

Data collection oriented objects

- frames (frequently registers)
- frame elements
- frame procedures
- sampling procedures
- samples
- measurement instruments
- data sources, enumerators, contact persons
- contact attempts, interviews
- measurement procedures
- observations

Data preparation oriented objects

- code lists, value sets, classifications
- coding procedures, classification procedures (including rules)
- editing procedures (including rules)
- error types (non-response, measurement errors etc.)
- suspected errors
- action types
- executed actions

Observation register oriented objects

- observation registers
- final observation registers: production versions and long-term storage versions
- data matrices

[10] Detailed descriptions of these objects are found in other parts of this Manual.

Statistics processing oriented objects

- estimation procedures for statistical parameters
- estimates
- estimation procedures for the accuracy/uncertainty in the estimates of statistical parameters
- accuracy/uncertainty estimates

Output oriented objects

- dissemination objects (paper or electronic publications, databases, data files etc.)
- dissemination procedures
- users
- uses

2.1 What is a statistical survey?

As described in another part of the Manual, Statistics Sweden performs many kinds of statistical surveys.[11] Some are (cf. Figure 3)

- censuses and sample surveys
- ad-hoc surveys and repeated surveys
- situation-based and activity-based surveys
- cross-sectional and longitudinal surveys
- primary surveys (data collected directly) and secondary surveys (e.g. surveys based on available records, and surveys such as National Accounts, which are based on the results of other surveys)
- surveys to maintain/update registers

In this Manual we regard all these types of "surveys" as statistical surveys in a broad sense, and the SCBDOK template is intended for use in all existing statistical surveys. It should be noted that an actual survey frequently combines several of the characteristics in the list above. A repeated survey might e.g. both be cross-sectional and rely on a direct collection of data. A survey to update a register might also be a survey relying on available records (e.g. the Statistics Sweden Register of the Total Population is updated by civil registration records).

[11] We have chosen a very broad definition of the term "statistical survey" Our reason is that all these versions exhibit many common features, both theoretically and practically. It would be easy to disregard these common features, if the versions were regarded as different phenomena instead of as special cases of the same phenomenon.

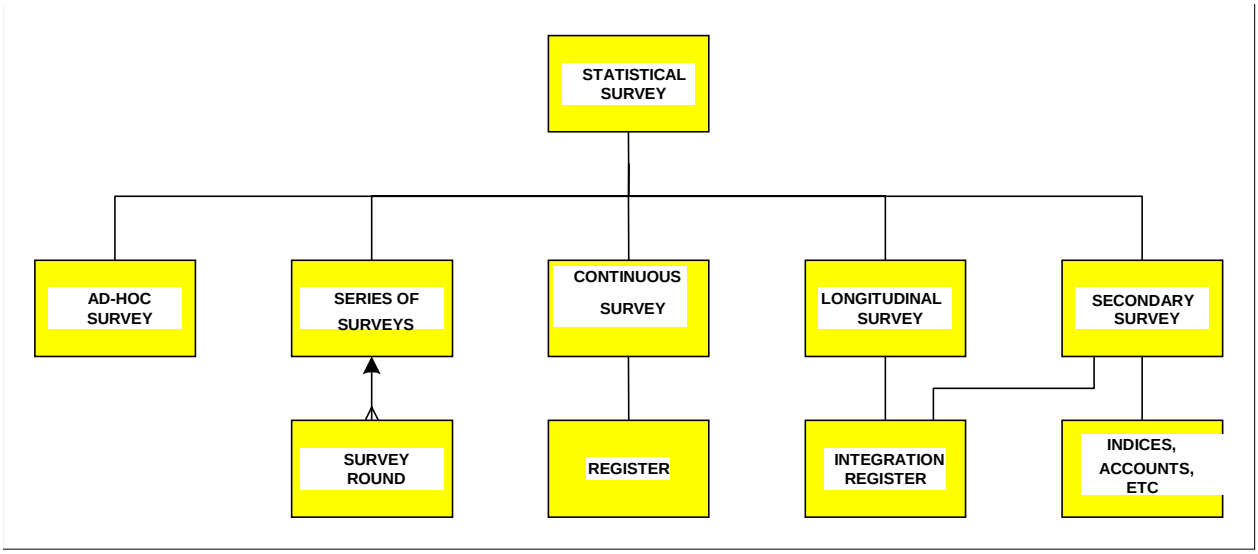


Figure 3. *Some versions of "statistical surveys" at Statistics Sweden.*

2.2 How to delimit a statistical survey?

The statistical surveys that constitute the basis for official statistics are usually repetitive, i.e. they are implemented at regular intervals, e.g. once a month, once a quarter, or once a year. This type of survey generates new final observation registers monthly, quarterly, and annually. Many aspects obviously remain the same from one round of the survey to the next, e.g. the main contents in the form of statistical parameters, while others, such as the magnitude of the non-response, change in each round. Each round and its final observation registers must consequently have its own documentation, even if its contents to a large extent overlap earlier and later rounds. It might be fairly minor changes from one round to the next, but on the other hand there is practically no survey design detail that isn't ever changed. In principle, all is changeable, although usually only a few changes occur at any one time. Some process data, though, are always unique in each round, e.g. the magnitude and composition of the non-response.

3 How to compile an SCBDOK documentation?

This Section provides some practical ideas on how to design the documentation work for a statistical survey according to SCBDOK. The following topics are covered:

- How to delimit *one* individual SCBDOK documentation
- How to draft the initial documentation
- How to document new rounds of the same survey
- How to harmonize SCBDOK documentations for various surveys
- How to document registers and continuous surveys
- How to document longitudinal registers and integration registers
- How to document surveys designed as systems of sub-surveys, e.g. CPI
- How to document national accounts and other secondary systems

- How to harmonize SCBDOK, METADOK and Quality Declarations
- The role of the electronic documentation tools

3.1 To delimit one individual SCBDOK documentation

An essential issue to consider before starting to draft an SCBDOK documentation is how to delimit it against other SCBDOK documentations. A simple rule seems to be that the documentation should cover *one* statistical survey, including the final observation registers that this survey yields. However, this simple rule has to be defined and supplemented in various respects.

The rule is easy to interpret when we talk about an ad-hoc survey of traditional textbook character. But most Statistics Sweden surveys are repetitive, and in addition they use data from (external) registers and other surveys together with the directly collected data.

In principle, a new SCBDOK document should be drawn up for each round of a repetitive survey. In other words, with *one* statistical survey (cf. the rule above) we mean one survey round. This seems to imply a lot of unnecessary duplication, as a repetitive survey usually remains fairly unchanged from one round to the next. Modern word processing (cut and paste), however, makes it simple to copy unchanged parts of a previous documentation. There is also an advantage in expressly having to decide, item by item, whether a survey really has remained unchanged or not. In some respects, each round is definitely unique, e.g. in the size of the non-response and the influence of other errors (process data). Each survey round also originates a unique final observation register that has to be documented and archived.

For surveys that are repeated in more or less identical form each month or quarter, it might, in spite of what has just been said, be useful to keep the survey rounds and their documentations together annually. The idea is to draft a common document, where each section first provides a general description that is valid for all the survey rounds during the year, and then goes on to indicate the deviations, if any, and the unique data (e.g. process data), for each individual round.

This approach is particularly suitable when the monthly or quarterly rounds form the basis for a corresponding annual survey.

3.2 To draft the initial documentation

During a start-up period, when we at Statistics Sweden have to make up for previous shortcomings in documentation, we unfortunately have to document surveys after they have been carried out, sometimes long afterwards. This implies various kinds of risks; even staff that actually participated in the survey work will not remember everything completely and correctly, some key person might have left, and there are new urgent tasks requiring our attention. But during a transition stage, until we have acquired improved documentation habits, we have no choice. We have to do the documentation afterwards.

Gradually we should build up a procedure where the initial documentation of a statistical survey is planned and designed in parallel with the survey. Figure 4 below, which will be discussed further in Section 2 of this Manual, illustrates the typical process of a statistical survey. Each step in the flow is treated both in the planning and in the implementation stages of the survey, and as this is done contributions to a SCBDOK documentation are generated spontaneously.[12] For instance, a discussion of the objectives and means of the survey will generate a list of the statistical parameters the survey is intended to estimate (Section 1.2 in the Manual). When we actually implement the various steps in the process, the process data are generated that are to be entered at appropriate places in the documentation, e.g. Section 3.3.

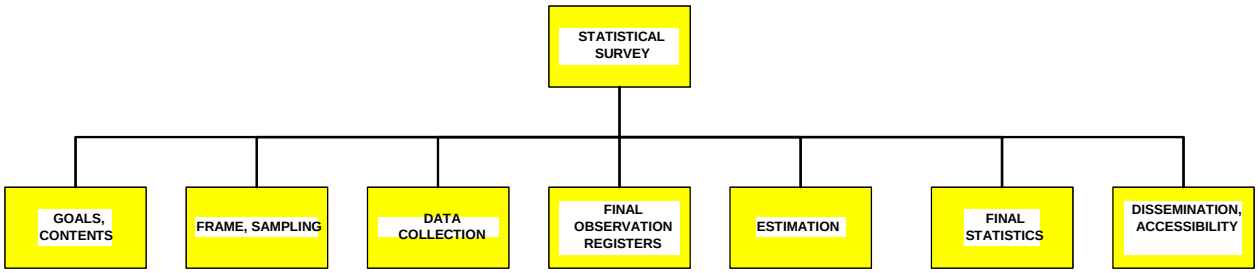


Figure 4. *Main flow of a statistical survey*

Consequently the planning, implementation and documentation of a statistical survey is synchronized in the following manner:

1 When the work to design the survey is started, the documentation of the survey is also started. As various decisions are taken, the corresponding sections in the SCBDOK template are filled in.
2 When the survey is carried out for the first time, the documentation is supplemented with the process data for this round, e.g. the frequencies of various types of error.
3 When the survey has been completed, the final observation registers and their documentation are frozen.
4 Then the documentation of the next survey round starts. A copy of the just finalized documentation is used and successively updated as in steps 1-3 above.

[12] The sections of this Manual adhere roughly to the chart of Figure 4. Section 1 defines the survey contents in terms of observation parameters and statistical target parameters. Sections 2.1-2.2 describe frame and sampling procedures. Sections 2.3-2.5 describe measurement instruments, collection procedures and data preparation. Section 3 describes the final observation registers. Section 4.1 describes estimation procedures, and Section 4.2 how the final statistics are presented, including procedures for dissemination and access.

3.3 To document successive rounds of the same survey

Item 4 above indicates the route. A copy of the final documentation of the most recently implemented survey round is used. Items 1-3 are repeated as the planning and implementation of the new survey round are once more carried out.

3.4 To harmonize SCBDOK documentations for various surveys

When a survey uses data from another survey or administrative records, the question emerges to what extent these data are to be documented. The main principle must be to ensure that these data are documented (according to SCBDOK) as carefully as the data that are collected directly from the respondents in the survey. If the data come from another SCB survey that has been documented according to SCBDOK, a reference (an electronic link) to that documentation is sufficient. If the data come from an external administrative register, we should at least obtain the appropriate administrative forms, including instructions, preferably in electronic form, and other essential information about the administrative processes.

A frequent issue refers to the Register of the Total Population, which serves both as a frame for many other surveys and provides the basis for some statistics. One alternative is to have one documentation that refers to its function as a frame, and another that refers to its function as a source of statistics.

3.5 To document registers and continuous surveys

How do we document a register that is updated more or less continuously? In such a case it is difficult to identify different versions. The same pertains to other event- and transaction-based statistical surveys, where new or updated data are flowing to the statistics producer more or less continuously.

In this case, the documentation problem is closely linked to the question of what constitutes a final observation register. For a register that is continuously updated, at least two types of final observation registers could be defined:

1 A transaction register that covers all transactions (events) reported before a specific last date, alternatively all events reported before the final date *and* which refer to events before an earlier date (e.g. 1 March or 1 January in the year x).
2 A situation register that reflects the actual register status in regard to population and variables at a specific date, alternatively the actual register status at a specific date, covering all transaction reported before a specific later date (e.g. 1 January or 1March in the year x).

3.6 To document longitudinal and integration registers

So-called longitudinal registers and integration registers also pose specific problems. A longitudinal register is an observation register that covers data from a survey series, i.e. a number of survey rounds. An integration register (which might also be a longitudinal

register) covers data from various surveys. Common to these types of observation registers is that not only are data from different surveys and survey rounds co-processed, efforts are also made to explain and remedy inconsistencies and other quality problems identified in the co-processings.

Longitudinal registers and integration registers require their own documentations, even if the registers employed are already documented. The documentation work might be harmonized analogously to what is mentioned above about harmonizing a number of monthly or quarterly surveys with a corresponding annual survey, which entirely or partly is based on the monthly/quarterly surveys.

3.7 To document surveys designed as systems of sub-surveys, e.g. CPI

Some statistical surveys, such as the consumer price index, are designed as systems of sub-surveys. Here there is a choice between documenting the component surveys separately and drafting a common documentation with appropriate additions for the various sub-surveys.

3.8 To document accounts and other secondary systems

Secondary statistics are statistical products which (almost entirely) are based on data from other surveys. The national accounts and other analytical systems are typical. SCBDOK can and should be used for such products as well. In such cases, the data collection consists in selecting and obtaining data from various existing data sources. Assuming the data to come from other Statistics Sweden surveys that are documented according to SCBDOK, references (links) are sufficient. Such an SCBDOK documentation will instead focus on the various statistical processings and analyses, i.e. Section 4 of SCBDOK.

3.9 To harmonize SCBDOK, METADOK and Quality Declarations

Figure 2 outlined the relations between the various documentation systems at Statistics Sweden. Section 1 (including Figure 1) aims to explain why different types of documentation are needed for different users and uses. In this context it is justified to ask:

- Are the various documentation systems harmonized to avoid duplication of work, and can documentation work already done be re-used in an optimal way?

There are some basic principles for meta-information systems. One is that the same information never is to be entered manually more than once. Another is that the entered metadata are to be structured in such a way that they easily can be re-used for new purposes by both human beings and computer software.

At Statistics Sweden we try to apply these principles, although the present situation is far from ideal. A fundamental aspect is that we use the same contextual and descriptive

system in all our various documentation and metadata systems, viz. the descriptive model presented in Part 2 of this Manual is also found in MIS 2001:1[13] and in the basic document on the descriptive model by Rosén and Sundgren.[14]

Let us take a look at the correspondence between the templates for SCBDOK and the Quality Declaration.

Figure 5 shows the SCBDOK 3.0 template, Figure 6 shows the Quality Declaration template (QD) according to MIS 2001:1, and Figure 7 indicates how the various parts of the templates relate to each other. Generally, we find

- The general contents part (parts 1.1-1.2 in SCBDOK and part 1 in QD) agrees very well. If the documentation has been done according to one template it can be more or less copied to the other. However, there is no object graph required in the QD template.
- Parts 2 and 4 in SCBDOK correspond largely to part 2 in QD. However, QD does not ask for an explicit, total presentation of the used estimation procedure (formulas etc.), but discusses the statistics from an accuracy perspective. As a discussion of accuracy should include quite a lot of information about estimation, and vice versa, the differences are not necessarily large. With some forethought the same description should in the main be satisfactory both in the SCBDOK and the QD (at least with some help of the word processor's cut and paste function).
- Finally, there is good agreement between a number of items describing such functions as presentation, dissemination and accessibility of the statistics and microdata, and concerning references to further documentation.

3.10 The role of electronic documentation tools

Finally some comments about the function of documentation tools. The great part of the work with an SCBDOK documentation is almost entirely independent of specific software tools. A word processor (with a cut and paste function and a simple drawing tool) will carry far. The tools developed in conjunction with SCBDOK, i.e. PCDOK and METADOK, ensure that some structures remain identical, and thus becomes re-usable in software contexts; cf. metadata-driven systems. The main message is that a considered and careful SCBDOK documentation is drafted, not how it is done and with the help of which tools.

[13] "Quality concepts and guidelines for quality declarations of official statistics", Meddelanden i samordningsfrågor för Sveriges officiella statistik, MIS 2001:1 Statistics Sweden. By Eva Elvers and Bengt Rosén.
[14] Rosén, Bengt and Sundgren, Bo: "Documentation for reuse of microdata from the surveys carried out by Statistics Sweden" 1991-06-28.

SCBDOK 3.0			
0	**General information**	**1**	**Contents outline**
0.1	Policy area	1.1	Observation parameters
0.2	Domain of interest	1.2	Statistical target parameters
0.3	Part of Official Statistics of Sweden	1.3	Output: statistics and microdata
0.4	Responsible person	1.4	Documentation and metadata
0.5	Producer		
0.6	Mandatory duty to supply data to the survey	**2**	**Data collection**
0.7	Confidentiality and processing rules for personal data		
0.8	Appraisal and disposal rules	2.1	Frame and frame procedures
0.9	EU regulations	2.2	Sampling procedures
0.10	Objectives and background	2.3	Measurement instruments
0.11	Use of the statistics	2.4	Data collection procedures
0.12	Design and implementation	2.5	Data preparation
0.13	Planned modifications in future surveys		
3	**Final observation registers**	**4**	**Statistical processing and presentation**
3.1	Production versions	4.1	Estimation: assumptions and calculation formulas
3.2	Long-term (archive, terminal) storage versions	4.2	Presentation procedures
3.3	Experiences from the latest survey round		
5	**Data processing system**	**6**	**Log file**

Figure 5. *Documentation template SCBDOK 3.0*

The Statistics Sweden Quality Declaration template according to MIS 2001:1			
0	**Introduction**	**2**	**Accuracy**
1	Contents	2.1	Overall accuracy
1.1	Statistical target characteristics	2.2	Sources of inaccuracy
1.1.1	Objects/Units and population	2.2.1	Sampling
1.1.2	Variables	2.2.2	Frame coverage
1.1.3	Statistical measures	2.2.3	Measurement
1.1.4	Study domains	2.2.4	Non-response
1.1.5	Reference times/periods	2.2.5	Data processing
		2.2.6	Model assumptions
1.2	Comprehensiveness		
		2.3	Presentation of accuracy measures
3	**Timeliness**	**4**	**Comparability and coherence**
3.1	Frequency	4.1	Comparability over time
3.2	Production time	4.2	Comparability between domains
3.3	Punctuality	4.3	Coherence with other statistics
5	**Availability and clarity**		
5.1	Dissemination forms		
5.2	Presentation		
5.3	Documentation		
5.4	Access to microdata		
5.5	Information services		

Figure 6. *Statistics Sweden Quality Declaration template according to MIS 2001:1*

SCBDOK vs QD template		
Items according to SCBDOK	**Items according to QD**	**Correspondence?**
1 Contents outline 1.1 Observation parameters 1.2 Statistical target parameters	**1 Contents** 1.1 Statistical target characteristics 1.1.1 Objects/Units and population 1.1.2 Variables 1.1.3 Statistical measures 1.1.4 Study domains 1.1.5 Reference times/periods	**2 Correspondence good**
	1.2 Comprehensiveness	No correspondence
2 Data collection 2.1 Frame and frame procedures 2.2 Sampling procedures 2.3 Measurement instruments 2.4 Data collection procedures 2.5 Data preparation **4 Statistical processing and presentation** 1.1 Estimation: assumptions and calculation formulas 4.2 Presentation procedures	**2 Accuracy** 2.1 Overall accuracy 2.2 Sources of inaccuracy 2.2.2 Frame coverage 2.2.1 Sampling 2.2.3 Measurement 2.2.4 Non-response 2.2.5 Data processing 2.2.6 Model assumptions 2.3 Presentation of accuracy measures	**Correspondence good, but partly different aspects emphasized**
0.12 Design and implementation	**3 Timeliness** 3.1 Frequency 3.2 Production time 3.3 Punctuality **4 Comparability and coherence** 4.1 Comparability over time 4.2 Comparability between domains 4.3 Coherence with other statistics	Some aspects covered by SCBDOK 0.12 SCBDOK provides some information to assess these items
1.3 Output: statistics and microdata 4.2 Presentation procedures 1.4 Documentation and metadata	**5 Availability and clarity** 5.1 Dissemination forms 5.2 Presentation 5.3 Access to microdata 5.4 Documentation 5.5 Information services	**Correspondence good**
3 Final observation registers 3.1 Production versions 3.2 Long-term storage versions 3.3 Experiences from the latest survey round		No correspondence
5 Data processing systems		No correspondence
6 Log file		No correspondence

Figure 7. *A comparison between the SCBDOK and QD templates*

To document statistical surveys, observation registers and statistical production systems

Manual to SCBDOK, Version 3.0

Part 2: Descriptive model

Bo Sundgren

2001-11-01

(not included)

To document statistical surveys, observation registers and statistical production systems

Manual to SCBDOK, Version 3.0

Part 3: Review of the documentation template

Bo Sundgren

2001-11-01

SCBDOK 3.0			
0	**General information**	**1**	**Contents outline**
0.1	Policy area	1.1	Observation parameters
0.2	Domain of interest	1.2	Statistical target parameters
0.3	Part of the Official Statistics of Sweden	1.3	Output, statistics and miicrodata
0.4	Responsible person	1.4	Documentation och metadata
0.5	Producer		
0.6	Mandatory duty to submit data to the survey	**2**	**Data collection**
0.7	Confidentiality and processing rules for personal data		
0.8	Appraisal and disposal rules	2.1	Frame and frame procedures
0.9	EU regulations	2.2	Sampling procedures
0.10	Objectives and background	2.3	Measurement instruments
0.11	Use of the statistics	2.4	Data collection procedures
0.12	Design and implementation	2.5	Data preparation
0.13	Planned changes in future surveys		
3	**Final observation registers**	**4**	**Statistical processing and presentation**
3.1	Production versions	4.1	Estimation: assumptions and calculation formulas
3.2	Long-term (archive, terminal) storage versions	4.2	Presentation procedures
3.3	Experiences from the latest survey round		
5	**Data processing system**	**6**	**Log files**

Part 3: Review of the documentation template SCBDOK 3.0

In Part 3 we walk through the template, section by section, item by item. This does *not* imply that the documentation work *has* to be drafted in the order outlined by the SCBDOK template. As we see it, the documentation work could be performed in fairly different ways, depending on which of the following three situations is at hand:

1. We are documenting a survey after the event, i.e. a survey that has already been implemented, perhaps rather long ago
2. We are documenting a new round of a survey for which there exists SCBDOK documentations for earlier survey rounds. In this situation it is preferable to copy the SCBDOK documentation of the previous round and use it as the starting point. Only modifications since the last round, or data that are always unique for each round, e.g. process data, need to be newly described or redrafted.
3. We are documenting an entirely new survey or a survey that is to be radically modified. In this situation, the documentation work should start simultaneously with the design and planning of the survey. As the various design decisions are taken, the corresponding parts of the template are completed. Figure 1 illustrates how the work may progress.

???TILLÄGG – JAG SJÄLV FÖRSTOD INTE RIKTIGT UPPLÄGGET I FRAMSTÄLLNINGEN FRÅN BÖRJAN OCH TROR DET VORE BRA MED EN FÖRKLARING I STIL MED STYCKET NEDAN. OK

For easy reference, we follow the order of the template and employ the same numbering. Sections 1.1 – 4.2 all start with statements or questions (framed) of what is to be covered in that part of the documentation. The text then indicates the objectives of the specific section/item, discusses some problems that the documenter always or frequently has to face in the drafting work, and outlines some practical approaches to the work.

0 General information

At Statistics Sweden, this Section is identical with the corresponding sections in the Product Description Template (section A, items A.1 – A.13). If it is desirable to have a physically complete documentation version, these items can be copied into the SCBDOK documentation, otherwise a reference (preferably an electronic link) suffices.

Sometimes the SCBDOK documentation concerns a survey or product that isn't delimited in a way that corresponds one-on-one with a product as defined by a Product Description. It then becomes necessary to describe how the SCBDOK documentation and the Product Description(s) correspond, e.g. by stating which SCBDOK documentations that correspond to a given Product Description, or which Product Descriptions that correspond to a given SCBDOK documentation.

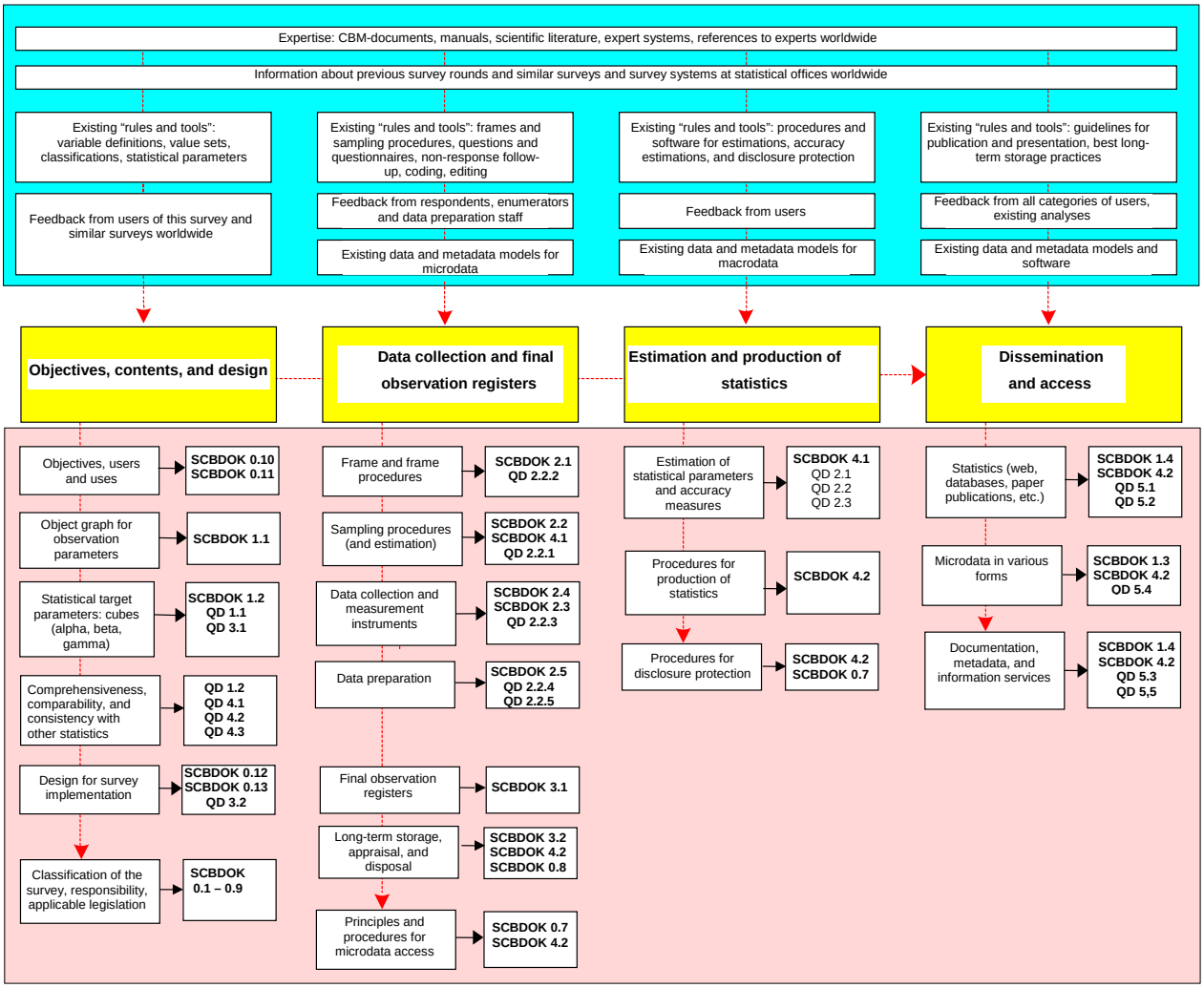


Figure 1. *How an SCBDOK documentation can be drafted in step with the planning of a statistical survey.*

1 Contents outline

This section of the documentation is to provide a general description of the main contents of the survey. The observation parameters are described generally in Item 1.1. The statistical parameters, which the survey aims to estimate, are outlined in Item 1.2. The output of the survey in the form of microdata and statistics is outlined in Item 1.3. Item 1.4 provides references to other documentation that describes various aspects of the survey.

Figure 1 is an object graph[15] that illustrates the most important documentation objects in Section 1 of SCBDOK and how they are related. It shows how a statistical target parameter is built up as a combination of a population, (distributed by) a number of gamma variables, (observed in respect of) a number of beta variables, (which are consolidated by means of) a statistical measure.[16] It also indicates that an observation parameter is a combination of a population and a variable.

[15] For a detailed description with several examples of object graphs, see Appendix 1 of this Manual.
[16] For a detailed description with examples of the alpha-beta-gamma structure, see Appendix 2 of this Manual.

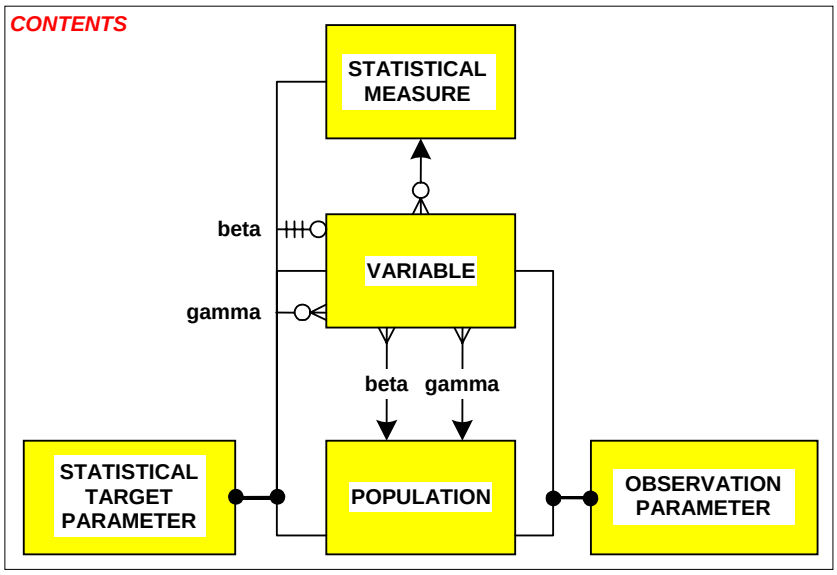


Figure 2. *Object graph showing the most important documentation objects in Section 1 of SCBDOK and their inter-relation to each other (cf. Figure 7 in Part 2 of the Manual).*

1.1 Observation parameters

- Indicate the observation parameters the survey collects or derive data from!

An object graph is a clear and satisfactory way of illustrating the most important observation parameters (see Figure 3). Further examples are given in Appendix 1 that also shows how to handle some typical modeling situations.

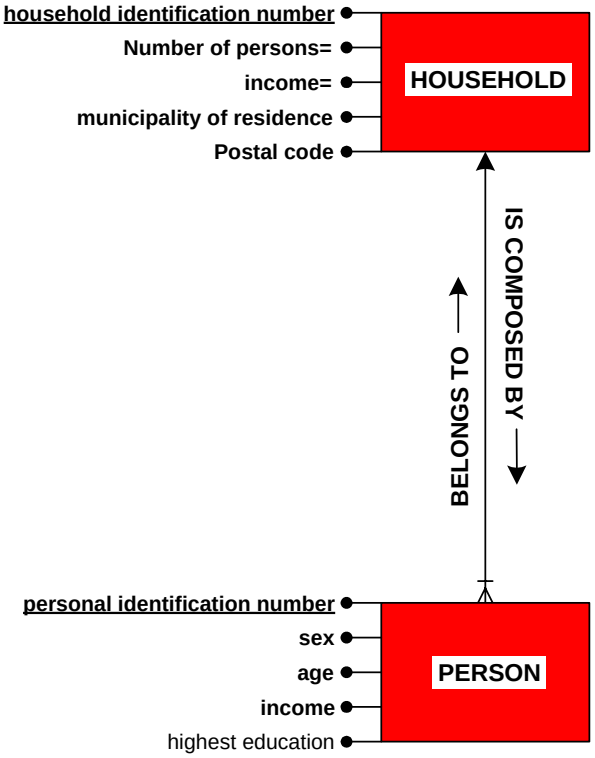


Figure 3. *Example of an object graph.*

1.2 Statistical target parameters

- Indicate the statistical target parameters the survey aims to estimate!

This item coincides with the item "Statistical target parameters" in the "Contents" part of the Quality Declaration.

A statistical parameter[17] is defined by

- the use of a specific **statistical measure** to summarize
- the values of a **variable** for
- the objects that are included in a specific **set of objects** (population/domain of interest)

Formally, we might then represent a statistical parameter with the triple

- O.V.f

where O is a set of objects, V is a variable, and f a statistical measure.

When we list the most important statistical parameters in the survey in this documentation item, it should be done systematically. The template below shows how this can be done:

[17] A more detailed discussion of the concept "statistical parameter" is found in Part 2 of the Manual.

STATISTICAL PARAMETER S = O.V.f: by the variables G	Parameter TIME(t) ι dimension	OBJECT SET O		SUMMARIZING MEASURE β dimension	
		POPULATION α dimension	DOMAIN γ dimension	VARIABLES V	STATISTICAL MEASURE f
S1: "Mean income during year t-1 of persons registered in Sweden at the end of year t: by municipality, sex and age."	Year t = 1995, 1996,…	Persons registered in Sweden at the end of t.	• Municipality(t): municipality where the person was registered at the end of t. • Sex(t): person's sex by the end of t • Age(t): person's age in entire years by the end of t	• Income(t-1): person's income during t-1 according to assessment in t	mean
S2: "Number of persons registered in Sweden by the end of year t: by sex, age and income interval."	Year t = 1995, 1996,…	Persons registered in Sweden at the end of t.	• Sex(t): see above • Age: see above • Income interval(t-1): person's income interval during t-1 according to classification xxx, based on the person's income during t-1.		number
S3: "Correlation between age at the end of year t and income during year t-1 of persons registered in Sweden at the end of year t: by municipality and sex."	Year t = 1995, 1996,…	Persons registered in Sweden at the end of t.	• Municipality(t): see above • Sex(t): see above	• Age(t): person's age in whole years at the end of t • Income(t): person's income during the year t-1.	correlation
S4: "National migrations during year t: by sex and income interval."	Year t = 1995, 1996,…	National migrations during t.	• Sex (t): sex of the moving person at the time of the move. • Income interval(t-1): income interval during t-1 of the moving person according to classification xxx, based on the person's income during t-1.		number
S5: "National migrations during year t from municipalities with higher tax rates to municipalities with lower tax rates: by sex and income interval."	Year t = 1995, 1996,…	National migrations during t, where the target municipality has lower tax rates than the municipality from which the person moved.	• Sex(t): see above • Income interval (t-1): see above		number
S6: "Per cent of national migrations in year t which went from municipalities with higher tax rates to municipalities with lower tax rates: by sex and income interval."	Year t = 1995, 1996,…	National migrations during t.	• Sex(t): see above • Income interval (t): see above	• To lower tax rates: move from municipality with higher tax rates in t to municipality with lower tax rates in t.	percentage Alternatively: S6 = 100%∗S5/S4

Figure 4a. *Template to specify the target parameters for a statistical survey.*

This type of structure is sometimes called a Tabulation Plan. It specifies in a systematic way the set of statistics that the survey primarily is to yield.

Figure 4b shows the object graph for a survey that might result in the statistics in Figure 4a.

Figure 4c shows two so-called star graphs, which illustrate the structure of the statistics in Figure 4a. In one of the star graphs (corresponding to the statistical parameters S1, S2, and S3 in Figure 4a) the focus is on the object type PERSON, while the other information is grouped around this object.

In the other star graph (corresponding to the statistical parameters S4, S5, and S6), the focus is on the object type MOVE, while the other information is grouped around this object.

Star graphs are often used to model multi-dimensional information in data warehouses.

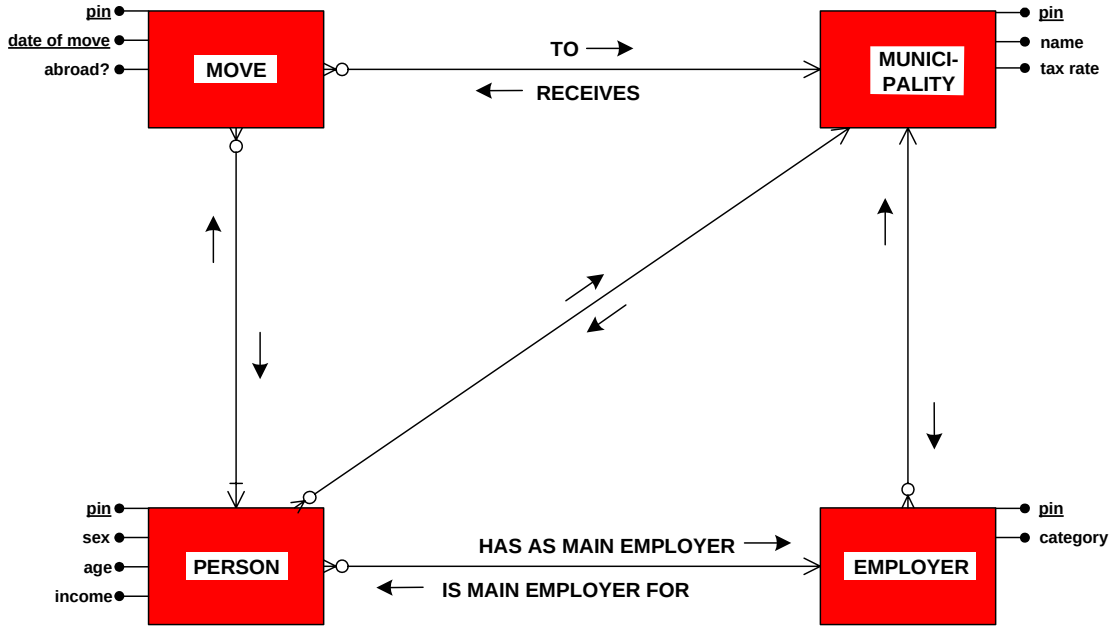


Figure 4b. *Object graph corresponding to the statistics in Figure 3a.*

1.3 Output: statistics and microdata

- Which observation registers are to be stored for further use?
- Which statistics are to be published? (Indicate dissemination form as well.)

Under this heading, the regular output of the survey is indicated, i.e. the data sets which in some form or other constitute the permanent results of the survey after the completion of a survey round. Primarily we mean

- Statistics published e.g. in the series Statistical Reports and/or in electronic form.
- Observation registers (microdata) which are long-term stored (alternatively in a database).

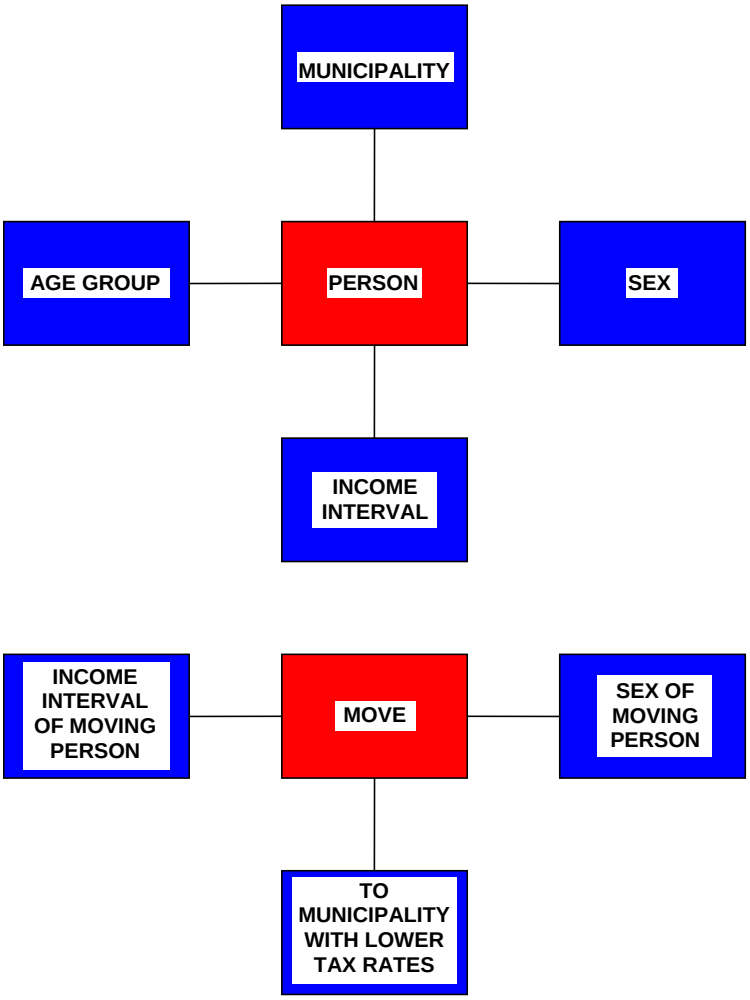


Figure 4c. *Star graphs to illustrate the statistics in Figure 3a.*

If Item 1.2 includes a Tabulation Plan, i.e. a systematic list of the statistical parameters that the survey aims to estimate, no further description of the contents of the statistics is necessary here; a list of the published statistics is satisfactory. For each publication, the publication mode (e.g. book, Statistical Report, CD-ROM, Internet, database) is to be ndicated as well as its name and unique identity code.

Similarly, the storage medium, title and unique identity code are to be reported for each preserved observation register (microdata). The contents and physical storage of the preserved observation registers are to be further described in Part 3.

1.4 Documentation and metadata

- List and describe the various kinds of documentations and metadata related to the survey (including process data) that are stored and accessible!

Under this heading we list other relevant documentation of the survey and its output, e.g. its Product Manual, Quality Declaration, and methodological reports. The

description should include the metadata (including process data) that are related to the survey and its output and are accessible in computerized form.

The references should preferably be supplemented by Internet and Intranet addresses, when the documents and metadata referred to are accessible in this way.

2 Data collection

Data collection is usually the most resource-demanding part of a statistical survey. Before the collection, a fairly detailed survey plan is usually designed that describes its various steps; frame procedure, sampling procedure (if a sample survey), the data collection proper, and the data preparation (coding, checking, correcting etc. of the collected data).

When we design the survey plan, we simultaneously get the structure of Part 2 of the SCBDOK. When the data collection then is implemented, there will sometimes appear disturbances that make it necessary to deviate from the plan in one way or another. We will assemble information about how the various sub-processes work (process data), e.g. experiences of how the measurement instrument functions and data about the magnitude and composition of the non-response. When the survey round has been completed, data of this kind are entered under Item 3.3 “Experiences from the latest survey round”. Such experiences may cause the survey plan to be modified in the next round.

A survey plan should provide answers to the following questions:

- *Who/What do we want to collect data about?*

The answer is the survey’s observation objects, i.e. the objects/units to be observed or measured in order to get the information about the target objects needed to estimate the values of the statistical target parameters.

- *Which data do we want to collect about the observation objects?*

The answer is the survey’s observation variables.

- *From whom do we want to collect the data?*

The answer indicates the survey’s data sources. If the source is a person, as when a person provides data about him/herself, we talk about respondents. But registers, databases etc. also serve as data sources.

- *How do we communicate with the data source?*

The answer describes the contact procedures. When the data source is a business or an institution, there is usually a designated person, who is responsible for submitting the data, their accuracy etc.

- *How do we carry out the actual observation (measurement) and the data entry?*

The answer describes the collection procedure, including rules for how to handle various kinds of disturbances in connection with the observation/measurement.

- *How do we prepare the collected data for the statistical processings?*

The answer describes the data preparation process. It includes such steps as data entry, coding, checking and correction. Frequently data preparation is to some extent an integrated part of the data collection.

- *How do we derive the target object and target variables from the observation objects and observation variables, when the latter are not identical with the former?*

The answer indicates the derivation procedures. We distinguish between primary data and derived data. The latter result when we, according to some rule, combine the primary data with each other or with other data available to the producer. An example of primary data is an answer obtained in an interview or income data obtained from the National Tax Board.

Note. For one and the same target population there might easily be more than one kind of observation object, and for one and the same observation object, data from various data sources might be collected.

Figure 5 indicates the most important documentation object in the data collection process.

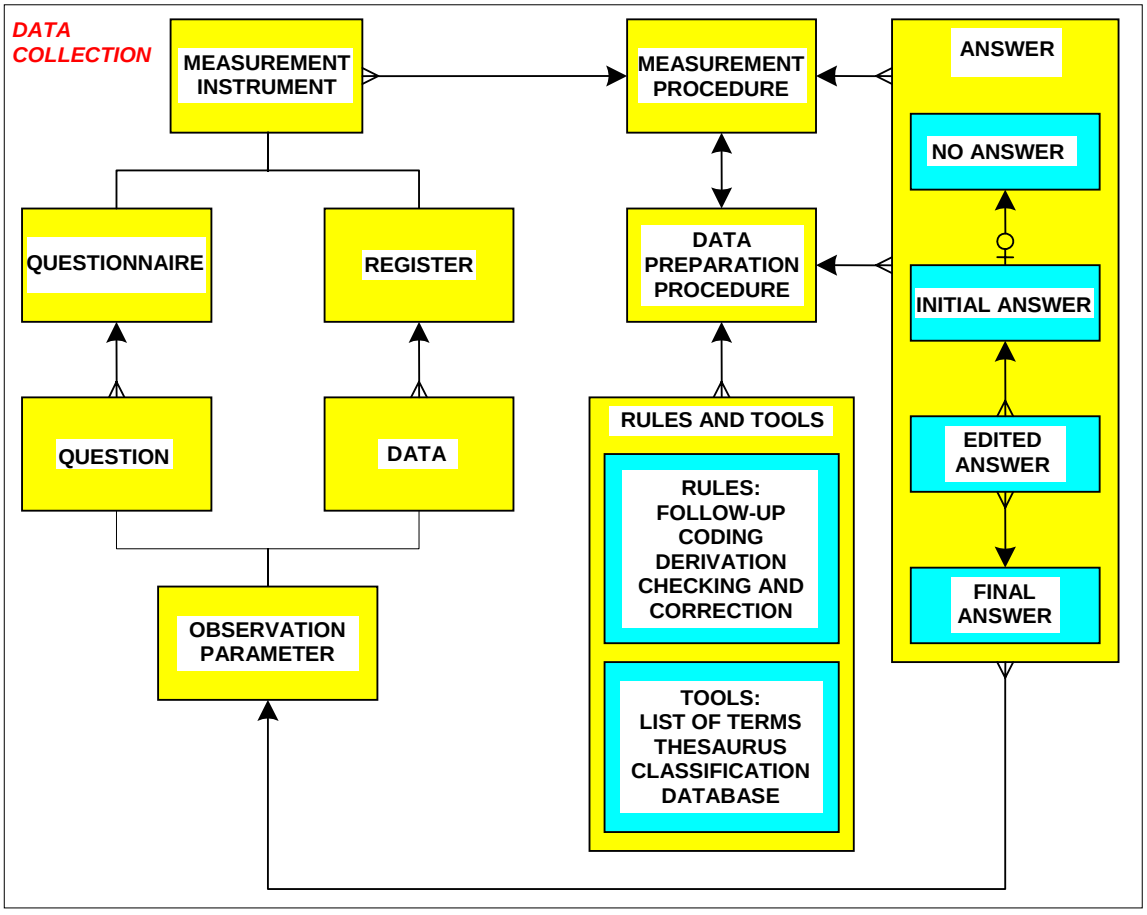


Figure 5. *Object graph that illustrates the most important documentation objects in the data collection process. (Cf. Figure 7 in Part 2 of the Manual.)*

2.1 Frame and frame procedures

- Describe the frame and the frame elements!
- Describe the links of the frame to data sources and observation objects!
- Describe the coverage of the frame; over- and under-coverage!

- Describe how the frame is established and maintained!

Under this heading we describe the frame and its links to observation objects and data sources. When appropriate, we also describe how we establish a frame adapted to the survey, and how it is maintained.

Figure 6 illustrates the most important documentation objects in the frame and the sampling procedures (Items 2.1 and 2.2).

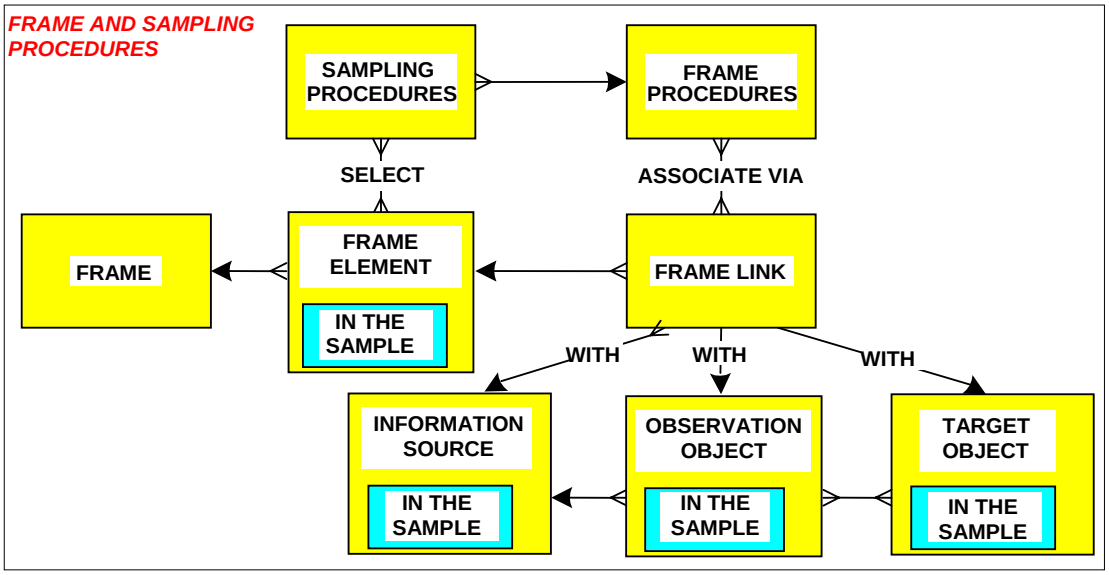


Figure 6.*Object graph illustrating the most important documentation objects in Items 2.1 and 2.2 about Frame, Frame procedures and Sampling procedures.(Cf. Figure 7 in Part 2 of the Manual.)*

The frame for a statistical survey consists of one or more lists that according to certain stated rules lead to data sources and observation objects. The objects listed in the frame are termed frame elements. “List” should be understood in a broad sense. Today the most common type of list is probably an electronic register, but it could also be e.g. a paper list or a map. Frequently there exists a frame, e.g. some administrative records, that can serve as the starting point. In other cases it might be necessary to create an entirely new frame for the specific needs of the survey.

The frame is a major facility for the data collection, irrespective of the type of survey.

The entire process to define and reach the observation objects and the data sources is called the frame procedure. The specification of the frame doesn’t in itself constitute the entire frame procedure. It also includes the identification of the links of the frame, i.e. the rules for how the frame elements are to lead to the observation objects and data sources. An important part of these procedures is consequently to provide the survey manager with a path to the data source (postal address, telephone number etc.).

In a simple case, the frame procedure for a sample survey may look like this: There exists a frame with elements that correspond one-on-one with the objects in the target population. A sample is drawn by random selection of a number of frame elements, and

the corresponding objects in the population constitute the sample. Data about the selected objects are collected directly from these objects.

In such a simple case there is a one-on-one correspondence between the

- frame elements
- data sources
- observation objects, and
- target objects.

In an actual case, two, three or even all four sets of objects may differ, and the correspondences between the objects in the various sets are not always one-on-one. This necessitates more complex frame procedures. In principle, there are two main types of data collections

- collection procedures based on the observation object
- collection procedures based on the data source.

In the first case, we first decide which observation objects we want. Then we look for data sources (might be the observation objects).

In the second case, we first identify some main data sources we want to utilize. We request these data sources to send us a list of all the objects (as specified) their register(s) cover(s) and their values on some specified variables. Only after we have received this information do we know exactly which observation objects that are covered by the source. The data collection for the identified objects may have to continue, and require us to turn to secondary/complementary data sources.

The collection models indicated above might be termed collection "by list" and "by cluster". It isn't always self-evident which of the models that is to be employed in a survey, and there are also surveys employing intermediary versions.

Frame coverage: under- and over-coverage

In a census (total enumeration), the aim is to collect the values of the target variables for all the objects in the target population, either directly or through derivation. The data collection in a census refers to all observation objects found in the frame. Ideally, we then get data for all objects in the target population. In practice, the situation is more complicated.

One complication concerns the coverage of the frame. The frame procedure might have resulted both in missing observation objects that ought to have been included, and in incorporating objects that ought to have been excluded. (In the discussion below we assume that observation objects and target objects coincide, which in principle they don't have to do.)

With frame population we mean the objects that the frame elements lead to via the links in the frame. The part of the frame population that falls within the target population we call the "attainable" part of the target population.

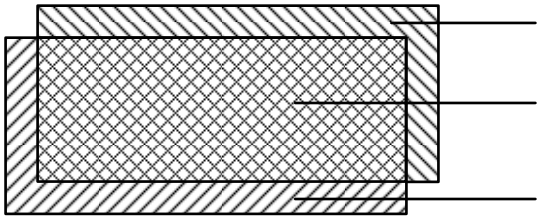


Figure 7. *Illustration of coverage.*

Objects in the target population that fall outside the frame population constitute under-coverage. They are objects that are not attainable through the frame procedure.

The part of the frame population that falls outside the target population is termed over-coverage. We assume that the links of the frame do *not* provide sufficient information to let us decide whether a frame element leads to an over-coverage object or not, and that such a decision would require actual observations. Or, put in another way, if the links do provide sufficient information on the object in question, the object can be excluded at once and will not constitute over-coverage. Figure 7 illustrates the concepts.

The under- and over-coverage concepts were introduced in the discussion on the assumption that the observation and target objects were of the same type, but the concepts can be used more generally. Under-coverage consists of objects in the target population that the frame procedure doesn't lead to data about; over-coverage consists of objects that the frame procedure leads to data about, although they are not part of the target population. Usually the collection of variable values for these objects is discontinued.

The consequences of under- and over-coverage differ generically. Under-coverage means that data are not collected for a part of the target population that the statistics are supposed to provide information about, and the consequence might be biased statistics. Over-coverage means that we devote survey resources to un-interesting observation objects. Here as well, if we don't check for population membership, the consequence might be biased (distorted) statistics.

Coverage problems appear in both sample surveys and censuses. On the assumption that observation and target objects are of the same type, the frame population for a sample survey is defined as the target objects identified by the sample frame and with a positive probability to be selected. With this modification, what has been said above about under- and over-coverage for censuses also pertains to sample surveys.

Frame maintenance

If the frame is to be used more than once, it has to be maintained and up-dated. This can be done by means of event-governed or inventory-governed frame updates.

In an event-governed frame update, a birth, death or other relevant change (e.g. change of address) more or less automatically initiates an update of the frame according to a fixed procedure.

In an inventory-governed frame update, the person responsible for the frame (e.g. the survey manager) periodically initiates a systematic up-date of the frame. There are also combinations of event- and inventory-governed updating procedures.

2.2 Sampling procedures

- Is the survey a census (complete enumeration) or a sample survey?

  If it is a sample survey, describe the sampling procedures:
- How is the frame stratified?
- Which sampling method is used (simple random sample, systematic sample, samples with differing sampling probabilities, cluster samples)?
- List sampling parameters such as stratum sizes and sample sizes!
- Which auxiliary information from the frame is used?
- Which algorithms and software are used?
- How are the actual values of the sampling parameters documented and stored?

A survey can be carried out either as a census or as a sample survey. In a census the ambition is to collect data for all the observation objects indicated by the frame.

In a sample survey, a sampling procedure is applied in order to use the allocated resources (including time) as rationally as possible. By means of the frame a sample of the observation objects/data sources are selected, i.e. we draw a sample of the observation objects/data sources the frame leads us to. The data collection in a sample survey pertains only to the sampled observation objects/data sources.

In sample surveys, we normally use probability samples. This implies access to a frame, the sample frame, that leads us to the observation objects/data sources. By means of a carefully controlled randomizing mechanism, we generate a sample of frame elements, the frame element sample. The randomization process has to be carefully structured to allow us to calculate the probability for any subset of the elements in the sample frame to be selected. The objects identified in the frame element sample form the (observation) object sample.

The reliability of the resulting statistics is influenced by the applied sample design and allocation pattern. Sample designs with stratified samples are common. The sample frame is then first structured into a number of different parts, (sampling) strata, from which sub-samples are then drawn independently of each other.

2.3 Measurement instruments

- Describe the measurement instruments used in the survey!

If some kind of questionnaire (including an electronic one) is used:

- Provide the entire questionnaire, including the full set of instructions. If the questionnaire and/or the instructions are very extensive, they can be given in an appendix.

Under this heading, we describe the survey measurement instruments, including the most important observation variables and their definitions. When a questionnaire is used, this is normally presented in its entirety.

When data are collected through the transfer of data from external registers or other data sources, we should describe the measurement instrument (e.g. the form) that was used in the original data collection, or, at a minimum, refer to some other readily accessible documentation of the data collection procedure and the used measurement instrument.

The concepts “measurement instrument” and “questionnaire” are discussed in more detail under Item 2.4 below.

2.4 Data collection procedures

- State the data collection methods used in the survey!
- Identify the survey’s data sources and describe the procedures to communicate with them!
- Describe briefly how the data collection proper is carried out!
- Describe discontinuation rules, if any, including measures at over-coverage!
- Indicate the checks implemented directly in connection with the data collection!
- Describe how non-response is counter-acted!
- Describe measures employed at object non-response and partial non-response! Describe substitution and imputation procedures, if any, including rules and tagging (coding) conventions!

Data collection methods

There are many data collection methods. One is quite simply to transfer data from one or more registers external to Statistics Sweden. Other methods are more direct, and we then talk in terms of measurements or observations, carried out by means of measurement methods and measurement instruments.

In surveys concerning individual persons, the questionnaire is the most common measurement instrument. Responses are obtained in mail surveys or in interviews. Interviews can be carried out as face-to-face interviews or as telephone interviews.

To attain a satisfactory measurement quality, the questionnaire form has to be well designed in respect to formulation, order of the questions etc. The construction of a questionnaire should include practical tests.

There are various technical solutions on how to measure. Some are personal computers with electronically stored questionnaires, touch-tone telephones linked to a computer, voice recognition designs, etc.

There are also physical measurement methods, used e.g. to provide data to environment statistics or crop yield statistics.

Data sources and communication routes
A data source is a body (person, organization, register) from which data are obtained. If the data source is a physical person, we prefer the term respondent. When the data source is a business enterprise or an institution, a person, termed contact person, is usually appointed to be responsible for providing the data, checking their accuracy etc. The procedures to establish contact with the data sources we call the survey's communication routes.

If the survey is implemented as a mail survey, procedures are needed

- to mail out questionnaires
- to tally received answers
- to mail out reminders.

If the survey is implemented by means of interviews (telephone or face-to-face interviews), procedures are needed

- to distribute the interviews on the enumerators
- to provide the enumerators with the means to communicate with the respondents
- to keep track of the interviews completed and remaining.

Discontinuation and measures at over-coverage
Discontinuation rules are needed when the frame procedure yields observation objects, from whom we don't want to start or continue the collection of data for the proper observation variables. The most common reason for this situation is that the frame includes over-coverage objects (cf. Section 2.1), and that this is discovered by means of some initial question. A discontinuation rule states that the data collection is discontinued as soon as we get certain (combinations of) answers to some questions.

Example. "If the answer to the last question is 'No', then stop and return the form."

Another reason for discontinuing might be that a certain value of some observation variable implies the values of other relevant variables.

Example. If the answer to a certain question implies that an enterprise is non-active, this implies zero values on other variables.

Non-response
Non-response occurs when a value for one or more of the observation variables for a selected observation object isn't obtained. Measures to counter-act non-response should be introduced at an early stage of the data collection.

If no usable data at all about a selected observation object are obtained, we talk about object non-response. The cause might be failure to establish contact ("no contact") or refusal to participate ("refusal").

If data are obtained for some but not all of the observation variables, we have partial non-response. Sometimes we can use imputation procedures, i.e. replace missing values with "probable values".

At object non-response, e.g. at failure to contact a selected observation object, we might choose to replace the object with another one according to some stated substitution rule. Such substitutions are uncommon in the surveys carried out by Statistics Sweden, but if they used, they should be mentioned under this documentation item.

The imputation and substitution procedures used should then be described, including the coding used to indicate missing and imputed values. Specific meta-variables can be defined to describe in detail the various kinds of non-response, implemented measures etc.

2.5 Data preparation

- Describe the rules and procedures for the various data preparation stages, viz. data entry, coding, checking and correction!
- Describe the derivation rules for the parameters derived from the primary data!

Note. To the extent that the data preparation measures are implemented as part of the data collection and are described in the previous section, the descriptions needn't be repeated here.

Here we describe the procedures prescribed by the survey plan for data preparation (coding, data entry, checking and correction).

- Data entry is often done together with the data collection. Sometimes a (secondary) entry procedure is needed, where data from forms that are not machine-readable are transferred to an electronic medium.
- Coding implies that data given in open-ended answers are replaced by codes from a specific code list according to stated coding rules. Coding is frequently done together with the data entry.
- Checking and correction can be done before, during and after the data entry. Both the checking and the correction should be carried out according to stated rules (to be included in the documentation).

Editing procedures can be classified in validity checks that are based on specifications of which valid values a given variable can assume, and logical checks.

The logical checks can be sub-divided into intra-object (intra-post) checks that check the consistency between various observation/measurement values for one specific object, and inter-object (inter-post) checks that check the consistency between the

observation/measurement values for different objects belonging to the same, or different, object types.

Macro-editing means the calculation of “provisional” statistics in order to identify suspected errors which significantly might influence the quality of the final statistics. The editing might lead to the discovery of (suspected) errors, which then can be checked and, if actually erroneous, corrected.

Derivation of parameters from collected primary data may also be considered as part of the data preparation stage.

Several of the processes listed above can be carried out

- as an integrated part of the data collection,
- and/or after the data collection proper has been terminated.

Combinations are common; the same type of data preparation measure can be implemented both in the data collection stage and at a later stage, and various types of measures can be carried out within the frame of one and the same process. Irrespective of order and combination, they are to be described in a suitable manner somewhere in the documentation.

3 Final observation registers

An observation register is a database for microdata. In it we store

- data collected about the observation objects
- derived data about the observation objects (and derived objects, if any, e.g. target objects of a different type than the observation objects)
- further data, if any, such as metadata, process data and/or other derived data, such as weights (see Section 4.1).

An observation register can have many different technical forms. It can consist of a number of flat files (or of tables) in a relational database, or the data can be stored in SuperCross or SAS databases.

An observation register passes through various stages, reflecting the various stages in the statistical survey. When we have finalized a survey’s target objects and target variables and its frame procedure (including observation object and observation variables), we have, at least in principle, outlined the planned observation register. It is to cover both the primary data collected and the variable values derived from these.

We can think of the planned observation register as a structure with “reserved positions” for the data we intend to collect. For each intended observation object (or derived object, e.g. target object) there is a post (“row”), and each post includes specific cells for each variable we intend to observe/measure (or derive) for the object in question. See Figure 8.

PERSONS in population P

PIN	Place of birth	Place of domicile	Income	Savings

Figure 8. *An observation register can be regarded as a structure with reserved cells for intended observation objects and derived objects, if any. Each post or row has specific cells for each variable it is intended to measure or derive for the object in question.*

As the data collection proceeds, the structure is filled by entering the observation values in the post and the cell that corresponds to its observation object and observation variable. We might add "status indicators" to the different cells, indicating whether a value has been entered; is or isn't expected to be entered; whether the value has been modified and in that case how and why; whether the value actually has been measured or is imputed etc.[18]

At some (defined) point in the life cycle of the survey round, we terminate the data collection and the data preparation, and in this way "freeze" the observation register. This final register then provides the basis for the production of statistics from the survey round as well as for database up-dates and future (re-)uses of the collected data.

Re-uses can take place both within and without Statistics Sweden. In the latter case, de-identified versions of the final observation register might be released to qualified users, e.g. university researchers.

With the same frequency as new survey rounds are implemented, new time versions of final observation registers are established. A monthly survey thus yields one final observation register each month, a quarterly survey one new register each quarter etc. If we draft a joint SCBDOK documentation for all the monthly surveys during one year, all the monthly versions of the final observation register must be documented in Section 3.

Sometimes the new time versions of the final observation register of a survey are accumulated in one joint final observation register for all the survey rounds implemented during e.g. a year. Alternatively, a successively incremented final observation register is established, covering all survey rounds since a certain point in time and forwards for as long as the survey is implemented. In the case of this kind of cumulated observation registers, it is advisable to freeze the register at least once a year and draft an up-dated documentation within the framework of that year's SCBDOK documentation. Both the frozen register and its documentation should also be long-term stored.

[18] These" status indicators" are actually examples of process data, e.g. metadata concerning the processes.

In addition to the various time versions of final observation registers, there might be other kinds of versions as well. In particular, it is important to distinguish between in-progress observation registers (production registers) and long-term stored registers.

As mentioned above, we should establish a final long-term storage (archive) version of the final observation registers yielded by a statistical survey during the year. Such an archive register has to satisfy the specific requirements defined by the National Swedish Archives. Briefly, the requirements state that it should be possible to read and interpret the registers (both in respect to technique and contents) in a distant future when the persons, computers and software of today have long since left the scene.

A typical case of an observation register-in-progress (production register) is a cumulated register of the kind described above, where new time versions of final observation registers are successively added. Such a final observation register is frequently structured as a database by means of some commercial software. The register will then be highly dependent on a specific technology and must be continuously maintained and up-dated by technically qualified staff in connection with e.g. change of software version or database manager. On the other hand, such a solution supplies good facilities to maintain a very high service level towards the statistics users, not least in respect to permitting fast and simple ad-hoc processings and analyses of interest to researchers.

To sum up, we normally have to manage a great number of different versions of final observation registers from a statistical survey:

- various time versions for various survey rounds
- various technical versions adapted to various software
- in-progress cumulative register versions (production versions), delimited in time or continually incremented
- long-term storage (archive, terminal) registers.

To actually structure and accommodate all these register versions and their metadata (including process data) requires quite a lot of thought. In respect to contents, there is usually substantial overlapping among the various versions of the final observation registers, even if the contents aren't necessarily identical. In technical respects, the various versions can differ substantially.

METADOK is a tool that helps to structure and order the documentation of the various versions of final observation and other registers from a statistical survey. METADOK also facilitates the documentation work itself in respect to basic contents and technical data.

METADOK thus helps to draft the greater part of the documentation needed in Section 3 of an SCBDOK documentation.

3.1 Production versions

This section should open with an object graph (see Section 1.1 and Appendix 1), to provide an outline of the observation and target area of the final observation registers.

The various registers, register versions etc. can then easily be related to the different parts of the object graph. In this way, a future re-user can quickly get an idea of the relevancy of the different observation registers to his/her problem.

The detailed contents and technical documentation of the observation registers is drafted by means of METADOK, and in accordance with its structure.

The object graph of Figure 9 describes this structure.[19] Detailed definitions of the concepts and of the relational data model by which the conceptual model has been implemented can be found in "METADOK 2.4 – Manual – Program to document formalized metadata for microdata".

3.2 Long-term (archive, terminal) versions

In this section we describe the registers or register versions which are to be long-termed stored in the same way as the in-progress registers discussed in the previous section.

[19] The structure is not entirely implemented in the presently existing version of METADOK. The shortcomings refer to the concepts "type of object" and "population".

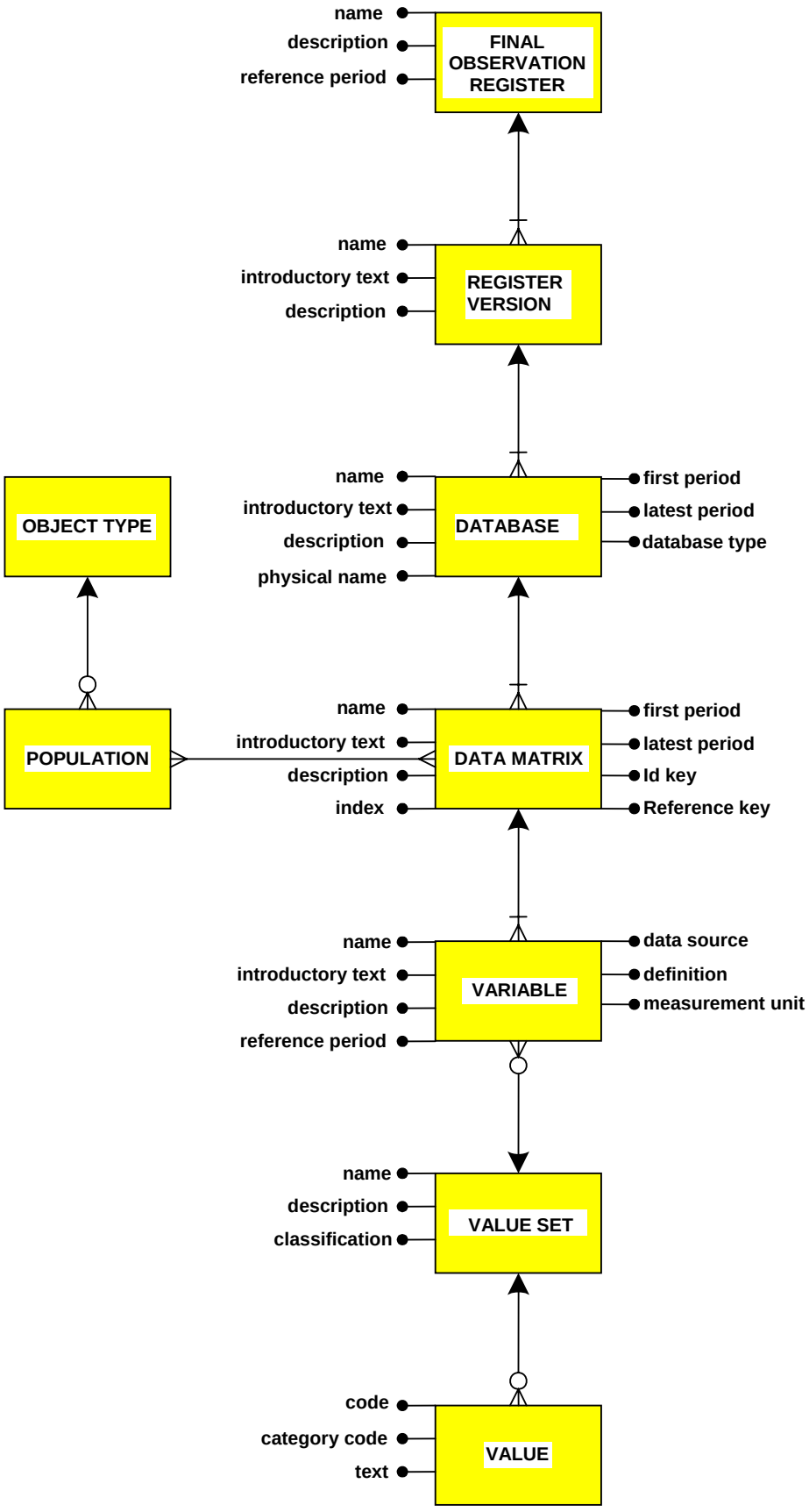


Figure 9. *METADOK conceptual structure.*

3.3 Experiences from the latest survey round

- Outline conditions and disturbances of various kinds that affected the just implemented survey round!
- How did the measurement tool and the collection procedure work? Comments from respondents and enumerators (if applicable)!
- Assess the accuracy (quality) of the collected data in various respects and totally! Are (some of) these estimates/assessments stored together with the final observation register or as part of it?
- Present the process information specific to this survey round, e.g. the time and cost of different production phases and the values of round-specific survey parameters such as stratum and sample sizes, non-response distributed by stratum, type of non-response etc! Are these data stored together with the final observation register?
- Edit/supplement the description of the final observation register!
- Will the survey plan (incl. the systems design) be modified in the next survey round?

Under this heading, we present information specific to the survey round which doesn't appear under the other documentation headings. In so far as the round has been implemented according to the survey plan, the information isn't to be repeated here. But it is important to present experiences, deviations from plan and their causes, the extent and consequences of different kinds of disturbances (e.g. under-coverage/over-coverage, non-response of different kinds, measurement errors, data entry errors, coding errors) and the actual values of important, round-specific parameters such as stratum and sample sizes.

It is desirable that as much as possible of this information - process data - is included as an integral part of the final observation register and becomes easily accessible to new users of the stored data. The same pertains to the structured metadata generated by METADOK (see Sections 3.1 and 3.2 above).

The final observation register is to be accompanied by a documentation showing both

- the survey plan (according to the previous documentation items) for the survey round that generated the observation register, and
- the round-specific information (according to this item) that modifies and supplements the survey plan information.

Disturbances and uncertainty
Practically always, statistical surveys suffer disturbances that prevent us from getting data in exact agreement with the survey plan. In turn, this means that the resulting statistics will suffer from more or less uncertainty, in addition to that implicit in using a sample. Some such uncertainty sources are discussed below.

We experience non-response when the values of one or more observation variables cannot be collected ("missing values"). Non-response also includes the case when data are missing because for some reason or other they couldn't be transferred from an administrative register.

If no usable data at all are obtained for an observation object, we talk about object non-response. Main causes are *unsuccessful contact (non-contact)* and *doesn't want to participate (refusal)*. At object non-response we might want to replace the non-response objects with others according to some stated substitution rule. (This is fairly uncommon in Sweden.)

If we get at least some usable data for an observation object, we term it a responding object. An object providing sufficient data to be identified as over-coverage is classified as responding. When an object provides usable data for some, but not all, observation variables, we talk about partial non-response.

In many cases a submitted particular does not agree with the "true" value in accordance with the variable definition. There are many reasons for this, e.g. that the question does not agree with the respondent's accounting system, that the question is ambiguously formulated, that the respondent miss-remembers, is careless (or worse, is consciously submitting false data), that physical measurement methods have shortcomings etc. The generic term is measurement errors. Measurement errors obviously contribute to the uncertainty of the statistics and they can do it in both a systematic and a random manner.

Below are given some examples of common types of disturbances and deviations from the survey plan, which should be reported under this heading. There are also some examples of other round-specific information that should be reported when applicable.

Frame generation

- Deviations, if any, from the frame procedure according to the survey plan (documentation Item 2.1);
- Problems, if any, incurred during the establishment and/or maintenance of the frame, and how they were managed;
- Quantitative data about the actual number of frame elements, frame up-dates etc;
- The actual distribution among various types of frame up-dates (event- or inventory-governed up-dates).

Sampling

- Deviations, if any, from the sampling procedure designed in the survey plan (documentation Item 2.2);
- Problems, if any, incurred in connection with the sampling procedure, and how they were managed;
- The actual values of some "sampling parameters" (stratum sizes, sample sizes), particularly if these haven't been stored as part of the documentation of the observation register (which they normally should be).

Data collection

- Deviations, if any, from the planned data collection procedure (documentation item 2.4);
- Problems incurred when trying to contact the data sources (e.g. in a mail survey or in interviews), and how they were managed;
- Deviations, if any, from the planned measurement procedure;
- Practical experiences from implementing the chosen measurement instruments (cf. documentation Item 2.3);

- Actual values of some process variables associated to the data collection process, e.g. time needed for interviews, number of contact attempts;
- Conditions attending discontinuation and over-coverage situations that have meant deviations from the survey plan or have created some kind of problem;
- Extent of non-response, more or less substantiated explanations of its causes, measures to counter-act it, and/or minimize its effects;
- Quantitative illustrations of the extent and distribution of the non-response. (Such data should be stored together with the observation register, to be easily accessible to re-users);
- References to specific non-response studies, if carried out;
- The actual results of the various data preparation phases, such as coding and checking, irrespective of whether these tasks have been carried out in association with the data collection or later;
- Description of how, in the production of the final observation register, the data on over-coverage, non-response, response, imputations etc. have been calculated and stored. This applies to the extent that additions and modifications are needed to supplement the survey plan. (According to "best survey procedures" these data should be accessible in electronic format together with the observation register proper, but it might still be advisable to emphasize important quantitative data of this kind in the documentation. Normally, the original observation register should be saved, including posts for which no answers were received.)

In a long-term perspective, as much as possible of the metadata and process data needed by a future re-user should be stored integrated with the data they describe. This can be achieved by means of "self-descriptive" file formats, such as GESMES.[20]

4 Statistical processing and presentation

The data in the final observation register constitute the basis for the statistics from the survey; primarily the statistics as specified by the Tabulation Plan (cf. documentation Items 1.2 and 1.3). The step from the final observation register to statistics is termed statistical processing and is described under this heading. The procedure for presenting the survey results is also described here, i.e. how the statistical results are produced, presented and analyzed.

[20] GESMES = Generic Statistical Message, an EDIFACT standard for transfer of statistical data and metadata. The standard is recommended by the European Central Bank (ECB) and Eurostat.

Figure 10 illustrates the most important documentation objects in Section 4 and their relation to each other.

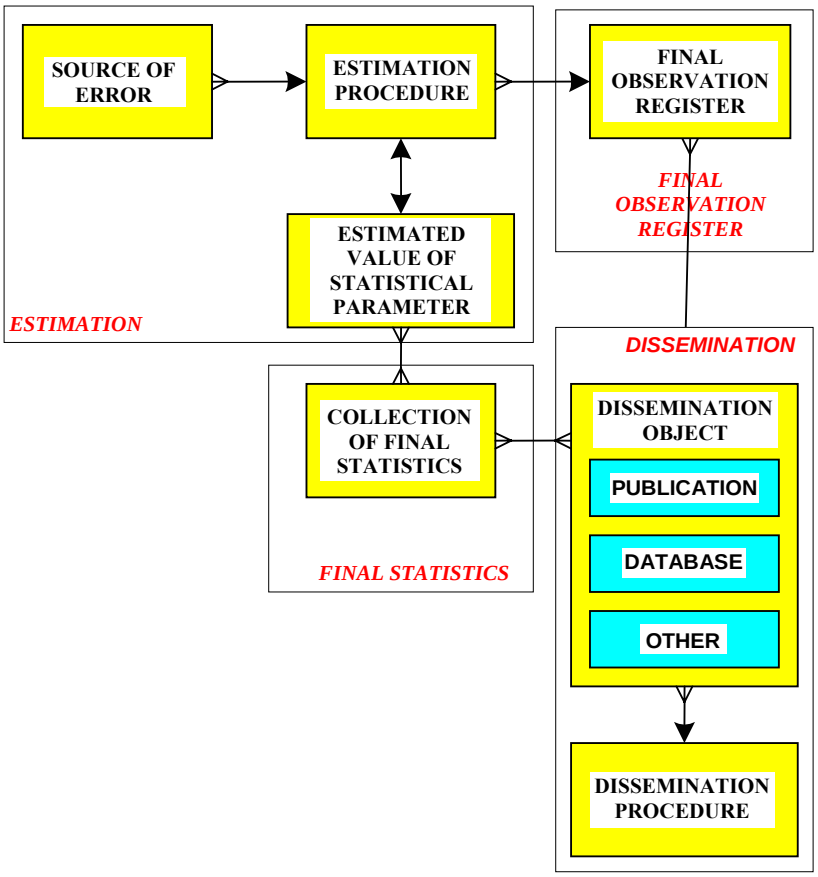


Figure 10.*Object graph that illustrates the most important documentation objects in the process "statistical processing and dissemination". Cf. Figure 7 in Part 2 of the Manual.*

4.1 Estimation: assumptions and calculation formulas

- Which point estimates and variance estimates are made, and which estimation procedures and calculation formulas are used?
- Which procedures are used for estimation/assessment of other quality characteristics than the sampling error?
- On which premises (observation models and population models) are the estimates based?

Point estimations

A (point) estimation procedure is a calculation procedure that brings us from observed values (primary data) to a statistic, i.e. an estimated value for a statistical parameter.

A common type of point estimation procedure in sample surveys implies that for each responding object, we calculate a weight, and that we then estimate totals by summing

the weighted observations (= *observed value x weight*). Weights are examples of derived variables that normally are saved in the final observation register.

In the estimation we must (explicitly or implicitly) consider the various kinds of disturbances that might exist, such as non-response, shortcomings in the frame, and measurement errors. Procedures with this aim we term adjustments, and they have to be based on e.g. models for how the disturbances arose. Non-response adjustment is also termed non-response compensation. A common way to adjust for non-response is to calculate the estimate as if the respondents constituted the original sample.

Statistical inference: observation models and population models

When moving from the final observation register to statistics, we usually have to face a statistical inference problem, as we normally don't know the exact variable values for all the objects in the target population that the statistics are to refer to. In a sample survey we always have to face an inference situation, but this is almost always the case in censuses as well, primarily because of coverage deficiencies, non-response and measurement errors.

In statistical inference, the conclusions, the estimates, have to be based on knowledge and assumptions about how observations and "reality" in the target population are related, observation models. Typically, these models are stochastic. They provide the basis for the estimation procedures, including various kinds of adjustments (e.g. for non-response, shortcomings in coverage).

The two basic principles for statistical inference are to use procedures leading to

- (at least approximately) unbiased estimates, i.e. estimates with little or no bias, and
- estimates with small estimator variance (standard deviation, mean standard error).

In sample surveys, our efforts to keep the estimator variance limited start with our choice of sample design and sample allocation. We then have to be guided by our knowledge and our assumptions about the situation in the population, for instance about how the values of some variables vary in the population. Assumptions about the conditions in the population are termed population models.

To minimize the estimator variances, we might use auxiliary information (easily accessible information about all the objects in the population or some totals). In this phase as well, we rely on population models.

"Population model" is an extensive concept, embracing models of different types. Some instances are:

- Assumptions about the variance of a variable in different strata;
- Homogeneity assumptions of the type "means at high aggregation levels are applicable at lower aggregation levels as well";
- Assumptions that a specific empirical time series can be computed according to a theoretical time series model.

An estimation procedure is said to be model *assisted* if the possible incorrectness of the model assumptions leaves the estimates unbiased; if not it is said to be model *dependent*.

Estimations of sampling errors (variance estimations)
Normally, a statistic is only an estimate of the statistical target parameter's (true) value. There is a deviation between the statistic and the value of the target parameter. In sample surveys, this is practically always the case.

As the true value of the target parameter is unknown, we cannot make any definite statements about the deviation between the statistic and the target parameter (if we could, we should eliminate the deviation).

Statements about deviations consequently become statements about the accuracy of the estimates. At best, we can talk about the probable highest value of the deviation. This is usually done by uncertainty intervals of the type:

- *uncertainty interval* = statistic ± *uncertainty margin.*

The interpretation of an uncertainty interval is that it with great probability includes the value of the target parameter value. Uncertainty intervals can be presented as confidence intervals, where the model premises can be formulated fairly objectively; or as assessment intervals??? based on subjective knowledge of the field of interest, survey experience etc. In official statistics, confidence intervals are most common. They are usually presented as

confidence interval = statistic ± *2 x (the estimated) standard deviation of the estimator*.

(2 is an approximation of 1.96 that is also used.) Generally speaking, such an interval has a confidence level of (approximately) 95%.

There are other indicators of the accuracy of statistics. Two of them are coefficient of variation and relative margin of error (= 2 [or 1.96] x the coefficient of variation). These uncertainty measures, like the confidence interval, are calculated by first estimating the variance of the point estimator. The calculation of uncertainty measures are therefore also referred to as variance estimation, and a calculation procedure for this purpose is termed a variance estimation procedure.

Estimations/Assessments of non-sampling errors
Ideally, an uncertainty measure is to refer to the total deviation between statistic and target parameter, but it is frequently difficult to provide such measures. Often it is easier to talk about the component deviations that relate to different error sources, such as non-response, measurement errors, coverage shortcomings, processing errors. In some cases, we also face uncertainty because we have used a model-dependent estimation procedure, and the model has turned out to be not quite correct.

4.2 Presentation procedures

- Which procedures are used to produce and publish the statistics?
- How are observation registers and statistics made accessible to users?
- How are observation registers and statistics analyzed?

On the condition that the survey's final observation registers are structured and stored in a commercially established, standardized way, most point estimates and variance estimates can be produced and presented in tables and graphs by means of standardized software.

Accessibility of statistics and observation registers
When the producer of the statistics has finalized an observation register and produced statistics and some analyses, the statistics and the observation register (within the limits of confidentiality and privacy regulations) are to be made available to the statistics users in appropriate media, such as paper reports, floppies, CD-ROMs, and database servers; and in appropriate channels, such monthly reports, annual reports, databases, Internet, Reuter. For Swedish official statistics, the dissemination of statistics is regulated in Acts, ordinances, directions and guidelines.[21]

Official statistics are to be published "as soon as finalized", in a printed paper report or electronically. The publication ought to be in accordance with a pre-determined time-schedule, and in a way that makes the statistics available to all users on the same conditions at the same time. The official statistics for which Statistics Sweden are responsible are to be published in Sweden's Official Statistics Databases.

When official statistics are to be published, the choice of publication forms, comprehensiveness of report, degree of detail and frequency of publication should be designed with regard to the demands of the users and their need for timeliness.

Swedish official statistics are considered publicly available by being available in the Statistics Sweden library and in some other libraries. All costs for production of official statistics, including the cost of publication and making them publicly available are covered by government grants. Costs for additional services, marginal costs, are to be paid by the users.

Official statistics are to be understandable to various user categories. Consequently, the statistics reports are to include descriptions and explanations as necessary, metadata. For official statistics there is also to be a Quality Declaration (QD), published and made available in the same way as the statistics in question. "Data materials providing the basis for statistics" are to be documented as well. The documentation should be of sufficient comprehensiveness and degree of detail to permit future use of stored data material, observation registers.

When official statistics are published electronically, a standardized format is to be used in order to permit storing the statistics in databases. Standardized formats are also

[21] The legal framework for Swedish official statistics is available at SAM-Forum, on the intranet of Statistics Sweden.

necessary to facilitate international reporting and the integration of statistics from various domains. The GESMES format (GESMES = Generic Statistical Message) mentioned earlier provides a suitable format. It is a self-descriptive format, i.e. a file stored in the GESMES format includes its own description (metadata and process data). GESMES has been developed as part of international statistical cooperation.

5 Data processing system

Directions under this heading are provided by the Statistics Sweden IT division.

6 Log file

In the log file we continuously enter all modifications we make in the production system (and consequently also enter in the documentation file). When we organize the observation register for a round in a survey series, we should up-date the log file in order to both

- get an overview of the most important modifications in the design of the survey which have been introduced between the different rounds and which influence comparability over time; and
- get a presentation that is complete in respect to the modifications in the survey design that have been introduced since the previous survey round.