Statistical Metainformation Systems

- *pragmatics, semantics, syntactics*

Bo Sundgren

R&D Report
Statistics Sweden
Research - Methods - Development
1992:17

INLEDNING

TILL

R & D report : research, methods, development / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1988-2004. – Nr. 1988:1-2004:2.
Häri ingår Abstracts : sammanfattningar av metodrapporter från SCB med egen numrering.

Föregångare:

Metodinformation : preliminär rapport från Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån. – 1984-1986. – Nr 1984:1-1986:8.

U/ADB / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1986-1987. – Nr E24-E26

R & D report : research, methods, development, U/STM / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1987. – Nr 29-41.

Efterföljare:

Research and development : methodology reports from Statistics Sweden. – Stockholm : Statistiska centralbyrån. – 2006-. – Nr 2006:1-.

R & D Report 1992:17. Statistical metainformation systems - pragmatics, semantics, and syntactics / Bo Sundgren.
Digitaliserad av Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-1992-X101OP9217

Statistical Metainformation Systems

- *pragmatics, semantics, syntactics*

Bo Sundgren

SCB

R&D Report
Statistics Sweden
Research - Methods - Development
1992:17

Från trycket December 1992
Producent Statistiska centralbyrån, utvecklingsavdelningen
Ansvarig utgivare Lars Lyberg
Förfrågningar Bo Sundgren, 08-783 41 48

ISSN 0283-8680
SCB-Tryck, Orebro 1993 01 Miljövänligt papper

STATISTICAL METAINFORMATION SYSTEMS

- pragmatics, semantics, syntactics

Bo Sundgren
Statistics Sweden
S-11581 STOCKHOLM

Invited paper for the Statistical Metainformation Systems Workshop
Luxemburg February 2 - 4 1993

Abstract

A conceptual framework is presented, covering pragmatical, semantical, and syntactical aspects of statistical metainformation systems. Examples from on-going projects in some statistical offices are used as illustrations.

0 Basic concepts

Put in a simple, but somewhat circular way, a **statistical metainformation system** is an information system, which informs about a statistical information system.

0.1 Statistical information systems

There are different kinds of statistical metainformation systems. One reason for this is that there are different kinds of statistical information systems. By tradition, most statistical information systems of statistical offices are **input-** and **production-oriented**. They are organized around **statistical surveys**, or **systems of statistical surveys**, which have related inputs and related production systems. Each survey is associated with a specific **data collection process**.

From a statistics user's point of view it is more appropriate to organize statistical information systems as **retrieval and dissemination systems** on the basis of the user's potential information needs. Such an **output-** and **user-oriented** statistical information system should provide a certain group of statistics users with a well organized, well integrated, and well described information potential, which is as relevant and complete as possible with regard to the needs of the users in focus.

Depending on the needs of the statistics users under consideration, a retrieval and dissemination system could be based upon one survey, a hand-full of surveys, or even all the surveys conducted by a statistical office, possibly in combination with a number of surveys and information systems, for which other organizations are responsible.

Thus, although it is, in principle, quite possible to regard each individual survey conducted by a statistical office as one statistical information system, it is usually more adequate, at least from a user-oriented perspective, to regard the individual surveys as subsystems of larger statistical information systems. Moreover, one and the same survey will often be a subsystem of more than one statistical information system.

0.2 Statistical metainformation systems

After this prelude, I will suggest a more precise definition of a statistical metainformation system, starting from a general definition of an information system.

Definition 1. An **information system** is a system, which helps a number of persons, the **users** of the information system, to establish and maintain their respective mental models, or **mind models**, of a certain piece of reality, the **object system**, or universe of discourse, of the information system. By performing this fundamental task, the information system can help its users to develop an **understanding** of the object system and its subsystems and components, and to plan, implement, monitor, and evaluate **actions** visavi the object system.

Definition 2. A **metainformation system** is an information system, whose object system is an information system, and a **statistical metainformation system** is an information system, whose object system is a statistical information system.

The following definition can now be derived as a corollary:

Definition 3. A **statistical metainformation system** is a system, which helps a number of persons, the **users** of the statistical metainformation system, to establish and maintain their respective **mind models** of a statistical information system. By performing this fundamental task, the statistical metainformation system can help its users to develop an **understanding** of the statistical information system and its subsystems and components, and to plan, implement, monitor, and evaluate **actions** visavi the statistical information system.

0.3 Global and local statistical metainformation systems

It is sometimes useful to make a distinction between global and local statistical metainformation systems. A **global statistical metainformation system** is a metainformation system, which informs about a complex statistical information system as a whole. A **local statistical metainformation system**, on the other hand, is a metainformation system, which informs about an individual survey.

A statistical metainformation system, which informs about a complex statistical information system, consisting of subsystems, sub-subsystems, etc, down to the individual surveys, will typically consist of a matching hierarchy of metainformation systems/subsystems, starting from a global metainformation system, passing through intermediate-level metainformation systems, and ending with a large number of local metainformation systems.

A global metainformation system is likely to be a relatively independent system in its own right, whereas a local metainformation system will typically be closely associated with, and often "built into", the individual survey production systems that it informs about.

0.4 Pragmatics, semantics, syntactics

Following Langefors (1966), Stamper (1973), and Malmborg (1989), I will distinguish between pragmatical, semantical, and syntactical aspects of information and information-related concepts.

The **pragmatical dimension** of information is symbolized by the question **"WHY?"**, and it concerns the **purposes** of information, that is, the relationship between the information and its **users** and **usages**.

The **semantical dimension**, symbolized by the question **"WHAT?"**, concerns the **meaning** of information, that is, the relationship between the *mind-internal* information as such and the *mind-external* **reality** that it refers to.

The **syntactical dimension** of information is symbolized by the question **"HOW?"**, and it concerns the relationship between, on the one hand, mind-internal information and information processes, and, on the other hand, mind-external (possibly computerized) **data representations** and (possibly computer-supported) **data processes**.

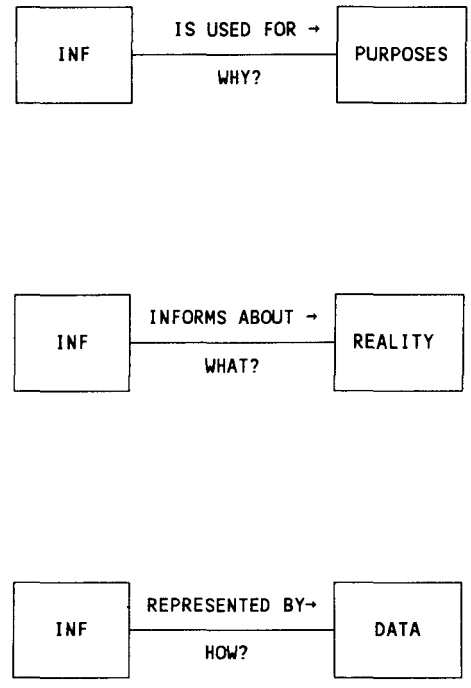


Figure 0.1. *Pragmatical, semantical, and syntactical aspects of information.*

This paper will give an overview of some important pragmatical, semantical, and syntactical aspects of statistical metainformation and statistical metainformation systems. For a more detailed treatment the reader is referred to Rosén & Sundgren (1991) and to Sundgren (1991a, 1991b, and 1992).

1 Pragmatical aspects of statistical metainformation systems: WHY are statistical metainformation systems needed? Who are the users, and which are their purposes and needs?

There are a number of different categories of users of a statistical metainformation system:

- "statistics users" users;
- "statistics producers" users;
- "designers" users; including
  - subject matter specialists;
  - statistical methodologists;
  - information system specialists;
- "managers" users;
- "software components" users.

Each one of these categories of metainformation/metadata users will have a certain typical **profile** of needs, and each profile will concern a typical combination of **semantical**, **syntactical**, and **pragmatical** aspects of a statistical information system and its component surveys.

1.1 "Statistics users" users

The most obvious purpose of a statistical metainformation system is to inform the users of the associated statistical information system about WHAT information they could obtain from the statistical information system, and HOW they can obtain it. This category of users will be referred to as the **"statistics users"** users.

Broadly speaking, the "statistics users" users need to

- search for, identify and locate possibly relevant statistical data;
- evaluate the meaning and quality of available data;
- judge how much time and money it would take to retrieve data;
- specify retrieval requests;
- actually carry out retrieval operations;
- interpret and process the results of retrieval operations.

A metainformation system, which supports both search and (full) retrieval operations, is called an **active metainformation system**. A metainformation system, which supports only search and (possibly) retrieval specification operations, is called a **passive metainformation system**. A passive system could help a statistics user to find relevant statistical data and (possibly) to specify a request for it, but it will not (itself) support the actual retrieval of data.

The "statistics users" users will typically need *a lot of* semantically oriented information, and *some* syntactically oriented information about the statistical information system and its contents.

A typical feature of the "statistics users" users is that they often combine statistical data (and other data as well) from several sources and production organizations. Thus they have a strong need for global metainformation.

1.2 "Statistics producers" users

Another purpose of a statistical metainformation system is to help those who operate the statistics production systems to remember what tasks they should perform, and how they should perform them. A related purpose is to train and introduce new staff in the production routines. This category of users will be referred to as the **"statistics producers"** users of a statistical metainformation system.

The "statistics producers" users (including low-level "managers" users of statistical metainformation systems; cf 1.4 below) will need very detailed knowledge about all aspects (semantical, syntactical, and pragmatical) of the surveys for which they are responsible. If they are experienced, they often know (or at least believe that they know) most of these details "by heart", and they may not feel a strong need for a formalized metainformation system, except possibly when they are engaged in maintenance and training activities. Neither do they usually feel a

strong need for global metainformation, unless they are confronted with very active and powerful statistics users.

1.3 "Designers" users

A third major purpose of a statistical metainformation system is to support different types of specialists, who design and maintain surveys and statistical information systems. This category of users will be referred to as the **"designers"** users of a statistical metainformation system, and it includes subject matter specialists, statistical methodologists, and information system specialists.

Each specialist category within the "designers" users group needs their typical profile of global and local metainformation. They need **global metainformation** in order to get hints and ideas from the design of "similar" surveys and information systems, and they need **local metainformation** about the particular survey or information system, which they are at present designing, redesigning, or maintaining.

Subject matter specialists have a particular need for pragmatically and semantically oriented details. **Statistical methodologists** focus on semantical and syntactical aspects of statistical procedures. **Information system specialists** analogously focus on semantical and syntactical aspects of databases and data processing procedures.

1.4 "Managers" users

A fourth possible purpose of a statistical metainformation system is to provide useful information to managers on different levels, who are responsible for the statistical information system as a whole, or some part of it. Among other things, these **"managers"** users of the metainformation system will need information about cost/revenue and quality aspects of the statistical information system, and about user attitudes and usage patterns visavi different parts of the statistical information and services provided by the system.

High-level "managers" users have a natural need for global metainformation, focusing on pragmatically oriented characteristics like user satisfaction, usage patterns, new demands, costs and revenues, timeliness, etc.

1.5 "Software components" users

Finally, a fifth purpose of a statistical metainformation system is to provide software components of the statistical information system (and the meta-information system itself) with **formalized metadata**, which are necessary, or at least useful, for running the software efficiently. We may refer to these "users" as the **"software components"** users of a statistical metainformation system.

The "software components" users typically require a lot of syntactically oriented metadata. More advanced and "intelligent" components will need a certain amount of semantically and pragmatically oriented metadata as well. For example, a **self-reorganizing database** will need data about usage patterns for different parts of the database, and an **expert system** supporting a design process or some kind of statistical analysis will certainly have to incorporate both semantically and syntactically oriented knowledge about the respective domains of competence.

1.6 The purposes of a particular metainformation system

A particular statistical metainformation system may have some or all of these purposes. Moreover, the different purposes may be more or less explicitly acknowledged. It is a **critical success factor** for a metainformation system that its designers have carefully considered what *should* be, and what *should not* be, a purpose of the system.

2 Semantical aspects of statistical metainformation systems: WHAT do statistical metainformation systems inform about?

2.1 Mind models of statistical information systems

According to *Definition 3* (section 0.2), a statistical metainformation system should help its users to establish and maintain their respective **mind models** of a statistical information system. Since the statistical information system is typically based upon a collection of statistical surveys, the statistical metainformation system should, *inter alia*, inform about these surveys.

More precisely, a statistical metainformation system should provide the information about its component surveys that its users need to perform their respective tasks. These tasks were discussed in section 1, and even though the discussion was very brief, it should have given the reader an idea of the kind of **mind models** of statistical surveys and statistical information systems that a statistical metainformation system needs to support.

If we are to specify the information contents of a statistical metainformation system, which should serve the needs of *several* or *all* of the above-mentioned user categories, we need to specify some kind of "ideal" or "standardized" **common mind model**, or **common description model**, a so-called **conceptual model** or **conceptual framework**.

2.2 A common description model for statistical information systems

Statistics Sweden has undertaken a project with the aim of creating a common **description model**, or **conceptual framework**, for statistical surveys and statistical information systems. Among other things, this work has resulted in a **documentation system**, called **SCBDOK**, based upon a standardized **documentation templet**, which is shown in *figure 2.1*. SCBDOK is now also being used as a basis for the design of a **metainformation infrastructure** for Statistics Sweden.

The information about a statistical survey requested by the documentation templet in *figure 2.1* is structured on the basis of the "natural flow of processes" in the planning and operation of a survey:

- *first* one specifies the **survey contents**;
- *then* a **survey plan** is developed;
- *then* **observations** are made, **data** are collected, prepared, and organized in some form of **database**, here called the **final observation register**, which is archived and possibly disseminated (in anonymized form);
- *then* the collected **microdata** are modelled statistically and transformed into **macrodata** by means of an **aggregation and estimation process**;
- *and finally* the resulting **statistics** are analyzed and reported through some suitable channel, for example a **publication** or a **statistical database**.

0 DOCUMENTATION STRUCTURE ETC

0.0 Documentation templet

0.1 Survey
- 0.1.1 Product number and product responsible person
- 0.1.2 System number and system responsible person
- 0.1.3 Statistics program and responsible person

0.2 Documentation modules and subsystems

0.3 Regularly reported or archived outputs

0.4 Related documentation

1 SURVEY CONTENTS

1.1 Universe of interest, verbal description

1.2 Universe of interest, formal description
- 1.2.1 Objects of interest
  - 1.2.1.1 Description
  - 1.2.1.2 Object graph
- 1.2.2 Populations of interest
- 1.2.3 Variables of interest

1.3 Survey outputs

2 SURVEY PLAN

2.1 Frame procedure
- 2.1.1 Overview
- 2.1.2 Frame and links to objects

2.2. Sampling procedure (if applicable)

2.3 Overcoverage/interruptions and undercoverage

2.4 Data collection procedure
- 2.4.1 Information sources and contact procedures
- 2.4.2 Measurement instruments

2.5 Planned observation register (incl derived concepts)
- 2.5.1 Overview
- 2.5.2 Object graph
- 2.5.3 Object/variable-matrixes
- 2.5.4 Definitions of derived concepts

3 DATA COLLECTION OPERATION

3.1 Sampling (if applicable)

3.2 Data collection
- 3.2.1 Communication with the information source
- 3.2.2 Measurement instrument
- 3.2.3 Data preparation at data collection time
- 3.2.4 Non-response, causes and actions
- 3.2.5 Substitutions (if applicable)

3.3 Data preparation (coding, data entry, editing, etc)

3.4 Production of the final observation register
- 3.4.1 Treatment of overcoverage/interruption objects
- 3.4.2 Treatment of non-response objects
- 3.4.3 Treatment of partial non-response
- 3.4.4 Counting of overcoverage, non-response, etc

3.5 Archiving and disseminating the microdata
- 3.5.1 Overview
- 3.5.2 File and record descriptions

4 STATISTICAL PROCESSING AND REPORTING

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

4.2 Estimation models

4.3 Computation formulas for estimation
- 4.3.1 Point estimations
- 4.3.2 Estimations of sampling errors
- 4.3.3 Judgements of other quality characteristics

4.4 Other inferences and analyses

4.5 Presentation and dissemination
- 4.5.1 *Printed outputs*
- 4.5.2 Electronical dissemination
- 4.5.3 Databases

5 DATA PROCESSING SYSTEM

5.0 System overview
- 5.0.1 Verbal description
- 5.0.2 System flow

5.1 Survey preparation (including sampling)
- 5.1.1 Overview
  - 5.1.1.1 Verbal description
  - 5.1.1.2 System flow
- 5.1.2 Component descriptions

5.2 Data collection operations
- 5.2.1 Overview
  - 5.2.1.1 Verbal description
  - 5.2.1.2 System flow
- 5.2.2 Component descriptions

5.3 Estimations, analyses, and reporting of results
- 5.3.1 Overview
  - 5.3.1.1 Verbal description
  - 5.3.1.2 System flow
- 5.3.2 Component descriptions

6 "LOG-BOOK"

Figure 2.1. *The SCBDOK documentation templet to be used for describing the surveys conducted by Statistics Sweden.*

If a statistical office has agreed upon a common description model for statistical surveys, and has managed to operationalize the description model in the form of a documentation system, like SCBDOK, it is a natural next step to introduce systematical working procedures for the development, operation, and maintenance of surveys. These procedures should of course be harmonized with the structure and contents of the description model and the documentation system, so that the different types of activities support each other, and so that the exchange of metadata is facilitated.

Figure 2.2 outlines a first version of a **systems development model** for Statistics Sweden, called **SCBMOD**. SCBMOD is harmonized with the documentation system SCBDOK.

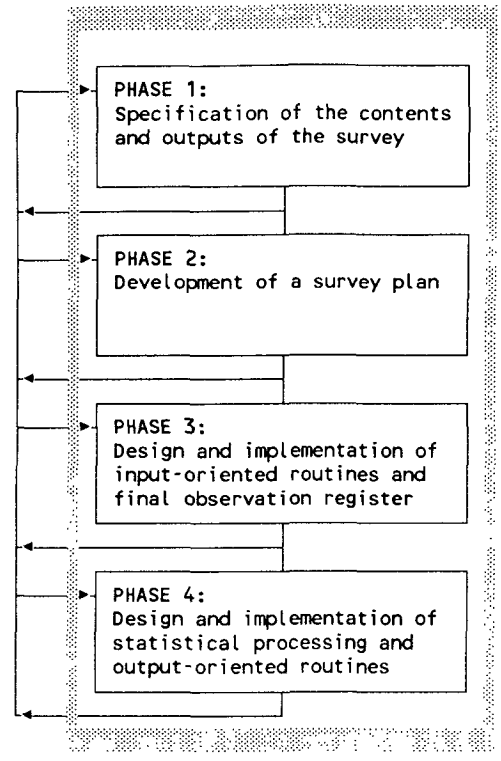


Figure 2.2. *Outline of a new systems development model, SCBMOD.*

An alternative to the flow-oriented way of structuring a specification of the information contents of a statistical metainformation system is to develop a more formalized conceptual model according to the **Object-Property-Relation-time** approach, **OPR(t)**; Sundgren (1973, 1974, 1984, 1989); or the similar **Entity-Attribute-Relationship (EAR)** model; Chen (1976), Elmasri & Navathe (1989).

Figure 2.3 shows an example of a so-called **object graph**, which is used for illustrating the structure and contents of a formalized conceptual model of an information system and its object system. Since we are now discussing metainformation systems, the graph is called a **metaobject graph**, containing **metaobjects**, **metavariables**, etc.

The **metaobject graph** in *figure 2.3* is a revised version of a metaobject graph showing a proposed conceptual structure of a **Data Catalogue** for the Australian Bureau of Statistics. The **metaobject types** indicated by small squares in *figure 2.3* should be interpreted as follows:

BOX "box structure" or "alfa-beta-gamma-tau structure" of macrodata; cf section 2.3.1 below;
POP population of objects (statistical units);
SAM sample of objects from a population;
XCL crossclassification of the population into (sub)domains of interest;
PAR parameter, statistical characteristic;
VAR variable;
VAS value set of one or more variables;
VAL value in value set;
SUR survey;

An asterisk at a place in the diagram, where a line from square A hits square B, indicates a "many"-relation, that is an object instance of type A could be related to more than one instance of type B.

In *figure 2.3* most (meta)object types occur in three versions: an **occurrence version** (*occ*), a **series version** (*ser*), and a **type version** (*typ*); corresponding to three layers of the conceptual model: an **occurrence layer**, a **series layer**, and a **type layer**. The division into three layers has the following background.

A typical pattern in statistical offices is that "the same" survey is repeated at regular time intervals, for example monthly, quarterly, or yearly. In such cases it is appropriate to speak about a **survey series**. Surveys producing indexes and other indicators (like unemployment rates) are typical examples of time series of "similar" surveys.

In reality, the different individual surveys within a survey series are never exactly identical; there are always some differences between the survey repetitions. It happens quite often that some component or aspect of the survey design is changed, if only marginally. For example, a new data item may be added, another one may be slightly redefined, etc. Even if the survey design should be exactly the same between survey repetitions, the conditions under which the survey is carried out will change, which will result in changes in response rates and other aspects of the quality of the survey data.

Thus the metadata for different survey repetitions within a survey series will be different, at least to a certain extent. *Both* the metadata generated by survey design decisions *and* the metadata generated by the survey process itself will change over time.

In principle, every item of metadata *may* change from one repetition of a survey to the next one. On the other hand, many relevant metadata items *will not* actually change between survey repetitions. A failure to recognize properly *both* the similarities *and* the dissimilarities between different survey repetitions in a survey series will negatively affect the **comparability in time**, an extremely important quality component for many users of statistics.

A similar problem concerns **comparability in space**, where "space" is a generic concept, covering not only geographical subdivisions, but also many other forms of classifications, where it is meaningful to recognize some kind of proximity and/or (fuzzy) similarity between different instances (occurrences) of one and the same type. For example, populations and variables with "similar" definitions may be good **substitutes** for each other with respect to certain needs.

The user needs for comparability in time and space must be taken into account when designing statistical metainformation systems. One way of doing this is indicated by the **three-layer model** in *figure 2.3*.

The **type layer** should contain metainformation, which is "usually" the same, or at least "similar" for different members of the same type. The type level meta-information has the character of **"general rules"** or **"typical descriptions"**; **exceptions** to the rules can be given for subtypes and/or occurrences of the types. This reminds of certain principles for knowledge representation used in artificial intelligence. It is also similar to the functioning of the human brain.

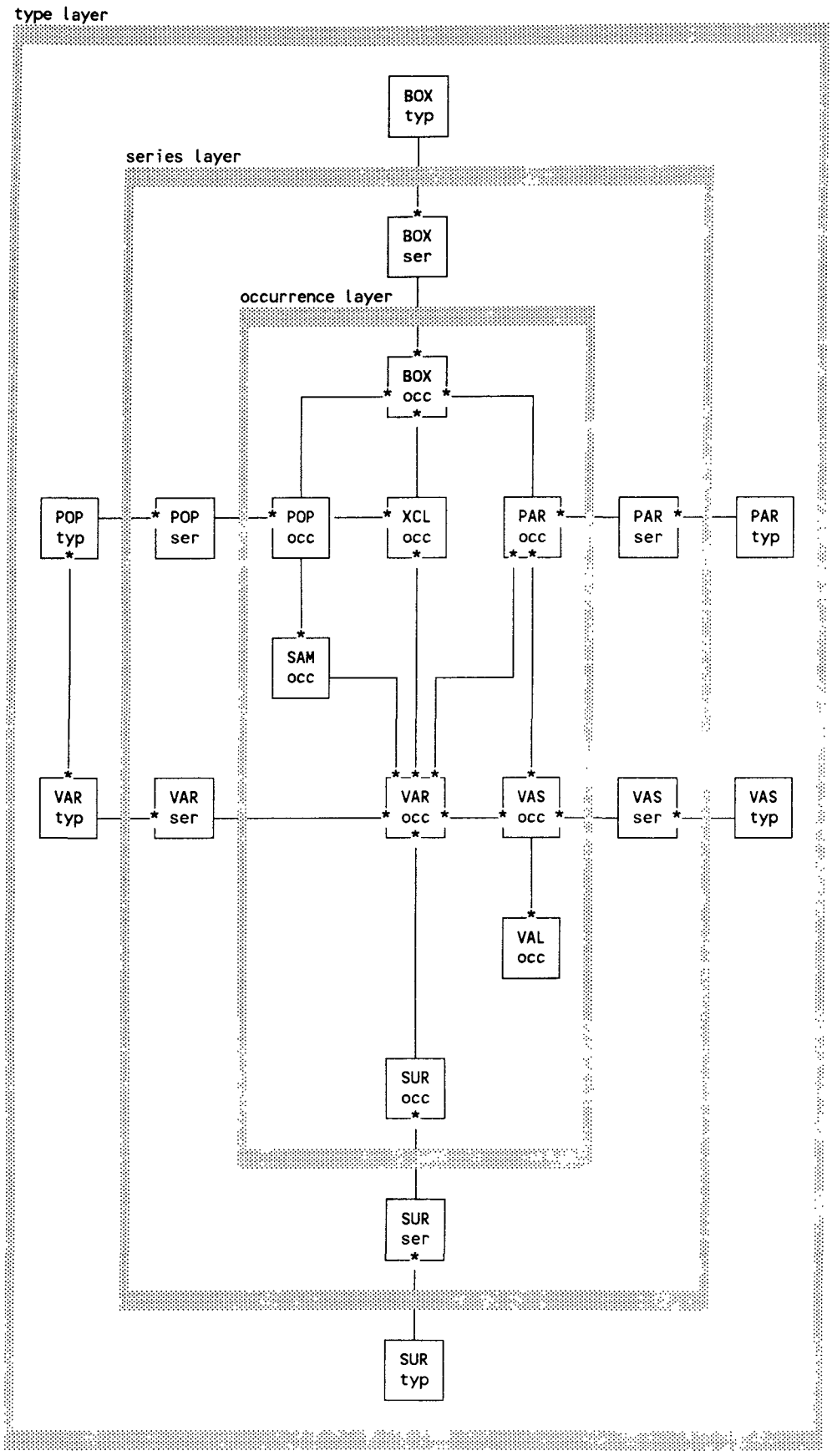


Figure 2.3. *A metaobject graph, which visualizes some aspects of the information contents of a statistical metainformation system. (The symbols are explained in the text.)*

Analogously, the **series layer** should contain metainformation, which is "more or less" the same for different repetitions within a time series. Once again exceptions to the typical descriptions can be given on the occurrence level.

The **occurrence layer** should primarily contain metainformation, which is known to be different, and maybe unsystematically so, between different occurrences within the same series, or the same type, respectively. High variability in this sense is typical for most **operation-based metavariables**, like "measurement problems" and "non-response rate". **Design-based metavariables** will not change their values between repetitions of "the same" survey to the same extent.

To summarize, most metavariables will have to be recorded on the occurrence level. However, if a metavariable is known to be relatively stable over time, it could be recorded on the series level, provided that there is an option to record **occurrence level exceptions** from the **series level rule**. The exceptions could result in **footnotes** in appropriate places, when the data are presented.

For example, if the measurement procedure for a variable is usually the same from survey repetition to survey repetition, the information about the measurement procedure could be given for the "VAR series" metaobject. If something unusal should occur with the measurement procedure during some particular repetition of the survey, this could be noted as an exception from the general rule, and the exceptional information would be recorded for the appropriate "VAR occurrence" metaobject.

If a metavariable is less stable, but still does not vary too much over time, it may be better to make the primary recordings on the occurrence level, but complement this information with some **"overview information"**, which is given on such a level of abstraction that it becomes stable over time.

For example, if response rates vary rather modestly over time, one could give information about the **"normal"** response rate span on the series level and give an "alarm signal" on the occurrence level, whenever the response rate falls outside the "normal span".

One could apply similar principles for determining the distribution of metadata between the type layer and the series layer of the metadatabase. "Normal" values of metavariables could be given on the type level, and exceptions from what is regarded as "normal" could be signalled on the series and occurrence levels.

2.3 Semantical aspects which are typical for statistics production

Many semantical aspects of statistical metainformation systems are similar to those of metainformation systems in general. I will not go further into such aspects here. Instead I will focus on the aggregation and sampling/estimation processes. These processes are typical for statistical information systems, and they have to be described rigorously by statistical metainformation systems.

2.3.1 The semantics of aggregation

The most typical feature of a statistical survey is that it contains an **aggregation process**, which transforms information about **individual objects** of a certain type, so-called **micro-information**, into information about **collectives of objects** of the same type, so-called **macro-information**. *Figure 2.4* explains the nature of this

process, which is so central and fundamental in all statistics production.

The upper part of *figure 2.4* illustrates the **object system level**, where we have a collective, O, of objects of the same type. Each individual object, o_i, in the collective is associated with a certain **true value**, x_i, of a certain **variable**, V. If we knew the true values of V for all objects in the collective, we would be able to compute the **true value** of a certain **parameter**, or **statistical characteristic**, P, for the object collective O, by means of a well-defined **aggregation function**.

In practice, we do not know all true values, x_i. What we have on the **information level** (cf the lower part of *figure 2.4*) is a set of **observed values**, v_j, for *some* (in the case of a **sample survey**) or *all* (in the case of a **complete enumeration**) of the objects in O. On the basis of the set of observed values we can obtain an **estimated value** of the parameter P for the object collective O, by means of an **estimation procedure**. The estimated value is interpreted as a substitute for the true value of P, and the **quality** of the substitute can be better or worse, depending on a number of circumstances.

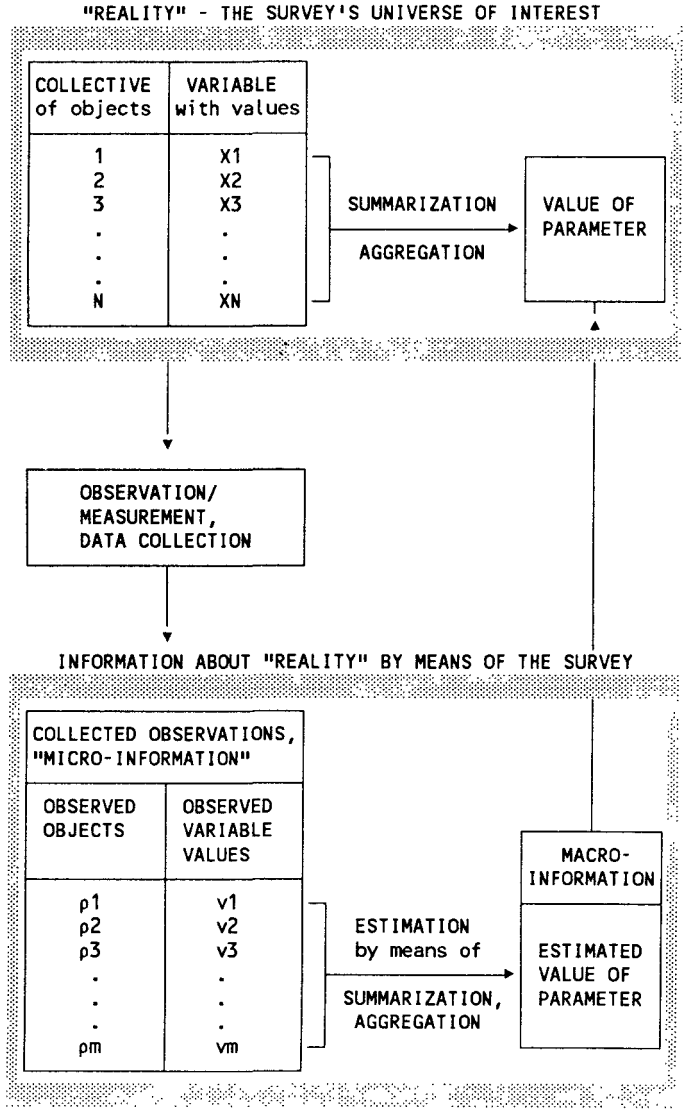


Figure 2.4. *Relationships between the universe of interest of a survey and the information about the universe of interest, which is observed, collected, and processed by means of the survey.*

An aggregation process results in **macro-information**, or **statistics**, which may be structured in terms of **statistical e-messages**. A statistical e-message consists of

- an **object component**, indicating

  - a **population of objects of interest**; which is sometimes
  - **restricted to a subset** by means of a **selective property**;
  and which is usually subdivided into
  - a set of **(sub)domains of objects of interest**; often by means of
  - a combination of **variables**; the value sets of which
  crossclassify the objects in the population;

- a **property component**, indicating

  - a **a parameter**, or **statistical characteristic**, which is estimated for the population as a whole, as well as for the domains of interest within the population; the parameter is usually defined in terms of
  - an **aggregation operator** (count, sum, average, correlation, etc) operating on one or more **aggregation arguments**, defined in terms of microlevel **variables** of the statistical units (objects, entities) in the population;

- a **time component**, indicating the (point or interval of) **time** at (during) which the population and its (sub)domains of interest existed and had the estimated parameter value.

The population part of the object component of statistical e-messages, including the selective property, if applicable, is referred to as the **alfa component of the statistical e-message**.

The crossclassification of the population into (sub)domains of interest is referred to as the **gamma component of the statistical e-message**.

The property component of statistical e-messages is referred to as the **beta component of the statistical e-message**.

The time component is referred to as the **tau component of the statistical e-message**.

Accordingly, the typical scheme of analysis for analyzing macro-information and macrodata is sometimes referred to as **alfa-beta-gamma-tau analysis**. *Figure 2.5* shows part of an example of such analysis from the Australian Bureau of Statistics; cf Sundgren (1991d). The structuring scheme has been applied to the statistical information published in the form of ordinary statistical tables in the August 1991 issue of *"Monthly Summary of Statistics Australia"*.

A multi-dimensional alfa-beta-gamma-tau structure is called a **box structure**; cf Sundgren (1973); or an **elementary abstract table (EAT)**; the latter term is used in Sundgren (1991d). It contains statistical e-messages with the same object component, but with different property components, and/or different time components. Thus an elementary abstract table will contain estimated values of one or more parameters at (during) one or more points (intervals/periods) of time for one set of domains of objects of interest within a certain population.

ALFA COMPONENTS	GAMMA COMPONENTS	BETA COMPONENTS	TAU COMPONENTS
Table 1 (page 1): Estimated resident population ('000).			
Persons resident in Australia at a certain point of time; *subset of "persons"*	State of residence	Count/1000	Yearly: 1985-06-30--1990-06-30 Quarterly: 1989-09-30--1990-12-31
Table 2 (page 1): Components of resident population growth, year ended 30 June 1990.			
Person events during a certain time interval, causing an increase or decrease of the number of residents in an Australian state; subset of "person events"	1. State of person event. 2. Event classification: birth/death, migration(overseas, interstate).	1. Population growth = sum of population growth contribution caused by event (+1 or -1). 2. Rate of growth = (1)/(number of resident persons at the ... of the time interval; *from table 1*).	The year 1989-07-01--1990-06-30.
Table 3 (page 1): Mean resident population ('000).			
Persons resident in Australia some time during a certain time interval; subset of "persons"	State of residence.	Mean resident population ('000) computed on the basis of counts for ... successive time periods according to the formula ...	One year periods: 1985, 1986, ..., 1990. Two year periods: 1984-85, 1985-86, ..., 1989-1990.
Table 4 (page 2): Live births registered.			
Live births registered during a certain time period; subset of "person events"	State of registration.	Count	Quarter years ending 1989-09-30--1990-12-31.
Table 5 (page 2): Deaths registered.			
Deaths registered during a certain time period; subset of "person events"	State of registration.	Count	Quarter years ending 1989-09-30--1990-12-31.
Table 6 (page 2): Marriages registered.			
Marriages registered during a certain time period; subset of "person events"	State of registration.	Count	Quarter years ending 1989-09-30--1990-12-d31.
Table 7 (page 2): Divorces granted.			
Divorces registered during a certain time period; subset of "person events"	State of registration.	Count	Years ending 1985-12-31--1990-12-31.

Figure 2.5. *Part of an alfa-beta-gamma-tau analysis of the tables in "Monthly Summary of Statistics Australia". Cf Sundgren (1991d).*

2.3.2 The semantics of sampling and estimation

In **sample surveys** sampling and estimation two closely related processes. *Figure 2.6* illustrates one possible way of modelling the semantics of sampled statistical information and of the processes of sampling and estimation, using some extensions to ordinary OPR(t) modelling; cf Sundgren (1989, 1991d).

The example used in *figure 2.6* is a sample survey, where the population is a set of object instances belonging to the object type PERSON. The values of some variables (*person#*, *region*, *category*) are assumed known for all the instances in the population. Parameters that are functions of these variables can be computed by evaluting the function over the object instances in the population. On the other hand *income* is a variable which is assumed to be relevant but not known for the object instances of the PERSON population. It should be estimated after observing a sample of PERSON objects. The sample is taken on the basis of random sampling from subsets of the population formed by stratification. Every object instance within a certain stratum has equal selection probability n/N, where n is the number of instances to be selected from the stratum, and N is the total number of instances in the stratum; n/N varies between strata.

The OPR(t)-model for the sample survey contains two object types corresponding to the object type PERSON: PERSON_IN_POPULATION and PERSON_IN_SAMPLE; there is a partial one-to-one relation between the two object types. The two other object types in the model, STRATUM and PERSON_GROUP_OF_INTEREST, can be formally defined as statistical aggregations of (any one of) the PERSON object types.

The formal definitions, expressed in the **infological language INFOL** (cf Sundgren (1989)) can be found in the text under the object graph. The meaning of the object type STRATUM is obvious from the name. The object type PERSON_GROUP_OF_INTEREST is an object type, whose instances are **domains of interest** or **domains of study**, that is, subgroups of the population (including the population as a whole) which are of particular interest for the users of statistics derived from the survey.

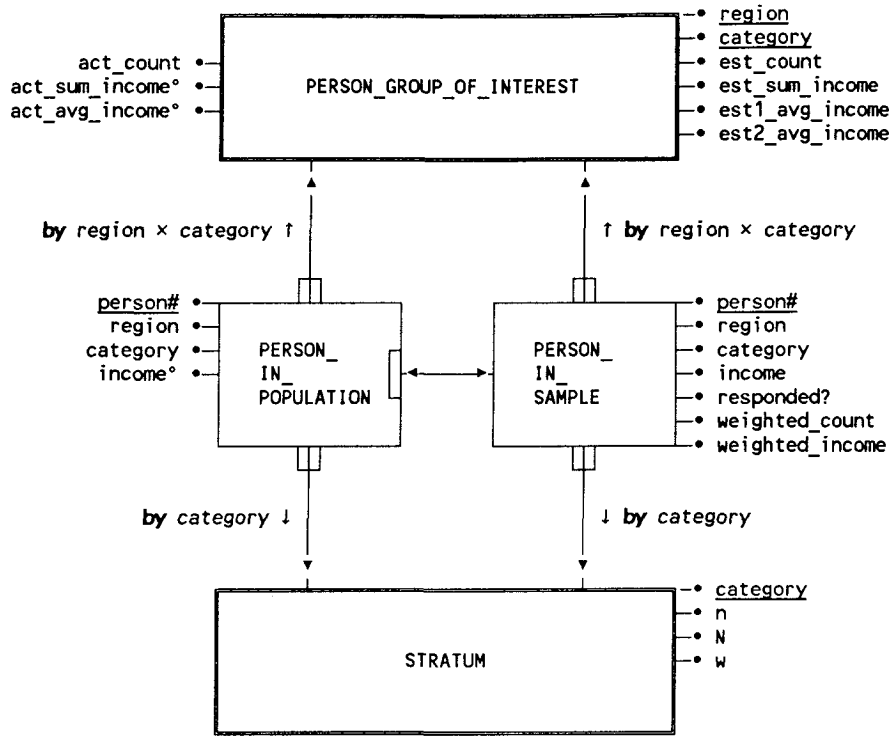
Many of the variables for the object types are derivable from other variables; once again the definitions are stated in INFOL below the object graph. Variables for which data are not available (like *income* for PERSON_IN_POPULATION) are indicated by a small circle (°) after the variable name.

3 Syntactical aspects of statistical metainformation systems: HOW do statistical metainformation systems perform their tasks?

There are two major syntactical problems associated with information systems in general, and with statistical metainformation systems in particular:

- How to organize the **data holdings**, which are used to store information in the information system? In the case of metainformation systems: *How to organize the metadata holdings?*

- How to organize the **subsystems, functions, and processes** in the information system, which are used to obtain, transform, and communicate information? In the case of metainformation systems: *How to organize the metadata processing subsystems, functions, and processes?*




```
Derivable object types:

PERSON_GROUP_OF_INTEREST <--- PERSON_IN_POPULATION(by region × category).agg;
PERSON_GROUP_OF_INTEREST <--- PERSON_IN_SAMPLE(by region × category).agg;
STRATUM <--- PERSON_IN_POPULATION(by category).agg;
STRATUM <--- PERSON_IN_SAMPLE(by category).agg;

Derivable variables for STRATUM:

n <--- PERSON_IN_SAMPLE(with responded="yes").count;
N <--- PERSON_IN_POPULATION.count;
w <--- N/n;

Derivable variables for PERSON_IN_SAMPLE:

weighted_count <--- STRATUM.w;
weighted_income <--- weighted_count * income;

Derivable actual variables for PERSON_GROUP_OF_INTEREST:

act_count <--- PERSON_IN_POPULATION.count;
act_sum_income° <--- PERSON_IN_POPULATION.sum(income°);
act_avg_income° <--- PERSON_IN_POPULATION.avg(income°);

Derivable estimated variables for PERSON_GROUP_OF_INTEREST:

est_count <--- PERSON_IN_SAMPLE.sum(weighted_count);
est_sum_income <--- PERSON_IN_SAMPLE.sum(weighted_income);
est1_avg <--- est_sum_income/est_count;
est2_avg <--- est_sum_income/act_count;
```

Figure 2.6. *An object graph - with accompanying INFOL definitions - corresponding to a sample survey.*

3.1 Metadata holdings

As was indicated by *figure 2.1* and *figure 2.3* above there is a need to store a lot of different kinds of metadata in a statistical metainformation system. The metadata can be categorized in several different dimensions, for example:

- by **metaobject type** (cf *figure 2.3*);
- by being **microdata-oriented** or **macrodata-oriented;**
- by **data type** (quantitative, qualitative, textual);
- by **type of formalism** (fixed-format facts, logical expressions, mathematical expressions, algorithms, graphs, free text);
- by being **information-oriented** or **process-oriented;**
- by being **procedural** or **declarative.**

Thus the metadata of a statistical metainformation system come in many different forms, and a relatively advanced **database management system** will be needed for handling the metadata holdings properly.

3.2 Metadata processing subsystems, functions, and processes

In a statistical office, every activity, which somehow manages data, should also manage the metadata, which is associated with the data.

In fact automation and computerization of survey management has up to recently implied **disintegration** of the natural relationships between statistical data and metadata, which existed in earlier manual systems. For example, consider a questionnaire. When it has been completed, it contains both data (answers to questions) and the associated metadata (the questions themselves and accompanying instructions for answering the questions). As long as the forms were processed manually, the data and metadata continued to go "hand in hand" throughout all the processing steps, until the final tables had been produced. Automation primarily aimed at rationalizing the counting process, a process which deals with the (object) data only. Thus the object data became separated from the metadata. When a programmer, in a later production step, was to compose readable tables, he or she would have to (re)introduce metadata, explaining the meaning of the data in the tables, but at that stage the original metadata (questions, instructions, etc) might very well have been lost track of. Thus the metadata in the presented tables would not normally be the result of a systematical, formalized transformation of the metadata in the questionnaires.

An essential feature of modern metadata management is that it is **reintegrated** with object data management, so that for example the metadata describing the figures in presented tables would in fact be the result of a chain of systematical, formally well-defined, and automated transformation processes, starting with the metadata in the questionnaire, or maybe even earlier, with the metadata generated by design decisions preceding the (computer-aided) construction of the questionnaire.

During all activities of all phases of the life-cycle of a statistical system, the different actors produce decisions, documents, etc, which contain metadata. If the metadata are properly captured and organized, they may become very useful, when the same statistical information system, or other ones, require metadata input.

It should be a challenge for every statistical office to organize its metadata processes in such a way that

- as many metadata as possible can be obtained from existing metadata holdings, whenever they are needed by a certain actor in a certain statistical system;

- as few metadata as possible have to be produced for its own sake, rather than as a side-effect of other (necessary) activities of the statistical systems monitored by the statistical office.

It follows that **sharing of metadata** (as well as sharing of object data) **within and between systems** should become a feature of rapidly growing importance for statistical offices aiming at rational, computer-supported planning and operation of its statistics production. A systematical, automated exchange of metadata between different activities in a statistical office promotes two good causes at the same time:

- it ***decreases*** **the burden** on those who would otherwize have to collect and enter the metadata manually;

- it ***increases*** **the benefits** from those metadata, which have already been collected and entered into computerized systems.

International standards for the storage and exchange of statistical data and metadata would significantly facilitate the efforts of all statistical offices to systematize and automate exchange of data and metadata, both internally and externally. Such standards will hopefully emanate from on-going UN/EDIFACT activities.

However, even before international standards have been established, statistical offices have very good reasons for streamlining their internal data and metadata flows. As regards metadata, useful work can be done in three directions:

1. **Create interfaces**, based upon preliminary, internal **standard formats** for the storage and exchange of (different kinds of) metadata.

2. **Look for possibilities to tap** useful metadata from manual, interactive, or fully automated processes, putting them into some kind of metadata holding or **metadatabase**, where they are stored in a standard format and are easily available for other useful purposes.

3. **Look for possibilities to feed** processes with existing (or automatically transformed) metadata from other processes or metadatabases, thus making the former processes (automatically) **metadata-driven.**

Figure 3.1 gives an example of a **metadata tapping and feeding procedure.** It indicates how some different components of a "total" documentation system of a statistical office could be coordinated, so as to minimize the manual metadata capturing work that has to be done. The basis for the "total" documentation is a so-called **production system documentation**, the primary purpose of which is to support the staff responsible for the operation and maintenance of the production system corresponding to a repetitive survey. The staff needs the production system documentation for such purposes as

- remembering the working routines between survey repetitions;

- finding out where and how to make changes in different components of the production system, when such changes are made necessary by changes in user requirements or other environmental conditions;

- training new staff members.

The production system documentation has to be updated at the same pace as changes are made in the production system. This implies a more or less "continuous" updating process. Whenever a change in the production system is made, the production system documentation should be accordingly updated, preferably in an automatical (or semi-automatical) way. In addition, a report about the change should be entered into a **log-book** in order to facilitate fast retrieval of all changes in a production system, which have taken place during a certain interval of time, for example, during the last five years.

A statistical survey typically produces two kinds of **end-products**, or results:

- **collections of observations (microdata)**, which are documented and archived for future reuse;

- **collections of statistics (macrodata)**, which are described and published via databases and/or traditional publications.

If the "total" documentation system is properly designed, most of the documentation needed for these two categories of end-products should be derivable as selected subsets, **"snap-shots"**, from the production system documentation, described above. Additional parts of the end-product documentations, which cannot be obtained by just copying some parts of the production system documentation, could anyhow be automatically obtained by means of formal transformations (derivations) on the basis of the production system documentation.

The production system documentation will typically reside with the organization responsible for the operation and maintenance of the survey. Thus the production system documentation will have the character of **local metadata**. The end-product documentations will typically follow the end-products, which means that they will often end up as parts of more **global metadata**. (Cf section 0.3.)

Figure 3.2 provides another example of metadata tapping and feeding. Most activities during the design of a production system for a survey will require, or at least benefit from, easy access to some metadata available somewhere in the statistical organization. For example, those responsible for designing the questionnaire of a new survey may considerably benefit from having easy access to questionnaires used for "similar" surveys in the past. When designing a new variable, it may be advisable to consult international standards. Etc.

Thus a design activity will ideally use a lot of metadata input. Some of these inputs will originally come from design decisions, which have been taken as the result of design activities in the past. Similarly, a current design activity will at some stage result in a design decision, implying certain metadata, which should be "tapped" from the design process, as automatically as possible, and "fed" into (primarily) a local metadatabase and (secondarily) more global metadatabases.

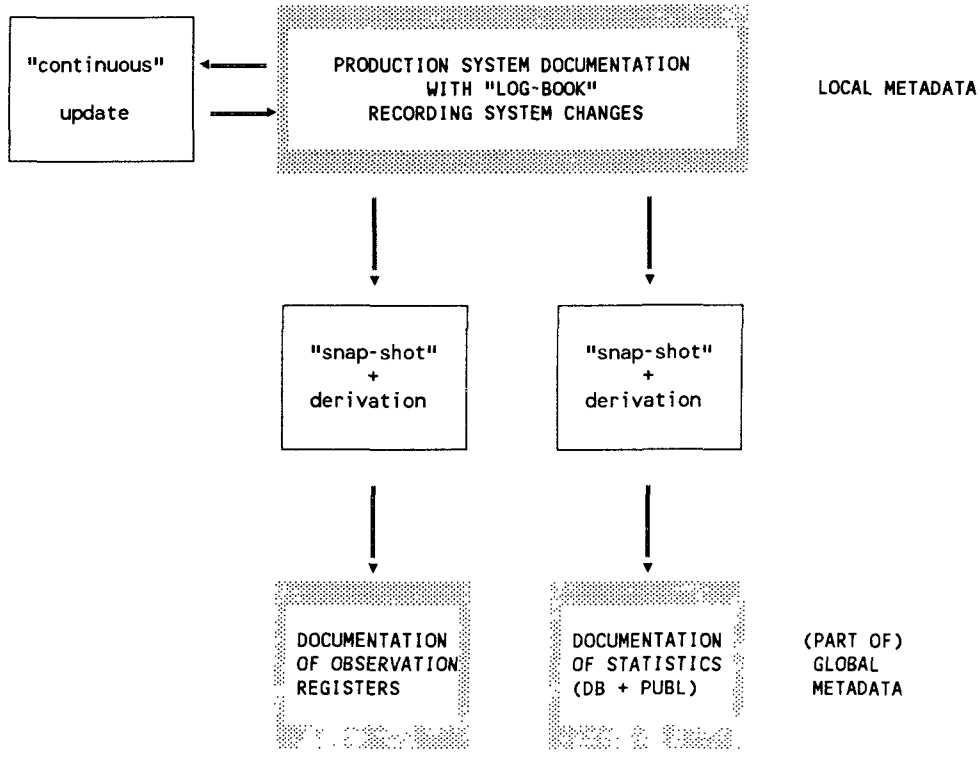


Figure 3.1. *Tapping metadata for survey end-products from production system metadata.*

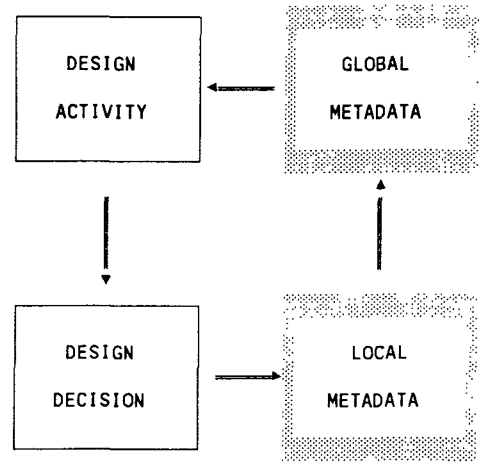


Figure 3.2. *Tapping and feeding of metadata between design activities of different surveys.*

Concluding remarks

Statistical offices are becoming increasingly aware of the necessity to provide - and provide efficiently - data and metadata, which are tailored to the needs of the users of statistics. This task is not quite simple, since statistical offices traditionally organize their survey production systems by input rather than by output, and different user categories have different, but overlapping needs for data and metadata. Data and metadata "are born", and must be captured, in the input-oriented structures, but then they must be "cycled and recycled" into a multitude of output-oriented structures. This report has suggested a number of concepts, principles, and models for tackling the problem in a way which economizes with human and other resources.

References

Chen P. (1976) "The Entity-Relationship Model - Toward a Unified View of Data" ACM Transactions on Database Systems, 1:1.

Elmasri R. and Navathe S. B. (1989) "Fundamentals of Database Systems" Benjamin/Cummings Publishing Company.

Klas, A. (1985) "The Metainformation System: Its Structure and Role in the Statistical Information System" Journal of Official Statistics Vol. 1 No. 4, pp 413-426.

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

Malmborg E. (1982) "The OPREM-approach - An extension of an OPR-approach to include dynamics and classification" Statistics Sweden.

Malmborg E. (1989) "On the Use of Semantic Models for Specifying Information Needs" Statistics Sweden.

Rosén, B. & Sundgren, B. (1991) "Documentation for reuse of microdata from the surveys carried out by Statistics Sweden" Statistics Sweden. Original report in Swedish. English translation available.

Stamper R. (1973) "Information in Business and Administrative Systems" London: B. T. Batsford.

Statistical Computing Project (1984) "Users Guide to Metainformation Systems in Statistical Offices" United Nations, Economic Commission for Europe, Conference of European Statisticians.

Sundgren B. (1973) "An Infological Approach to Data Bases" Statistics Sweden.

Sundgren B. (1974) "Conceptual Foundation of the Infological Approach to Data Bases." In Klimbie J. W. and Koffeman K. L. (eds) Data Base Management Amsterdam: North-Holland.

Sundgren, B. (1980) "Meta-Information in Statistical Agencies" Statistics Sweden.

Sundgren B. (1984) "Conceptual Design of Data Bases and Information Systems" Statistics Sweden.

Sundgren, B. (1989) "Conceptual Modelling as an Instrument for Formal Specification of Statistical Information Systems" ISI 47th Session, Paris.

Sundgren, B. (1991a) "What metainformation should accompany statistical macrodata?" Report for the June 1991 Meeting of Working Party 9 of the OECD Industrial Committee as a basis for a discussion on the topic of Standards for Metadata in International Databases. Also available from Statistics Sweden.

Sundgren, B. (1991b) "Statistical Metainformation and Metainformation Systems" Report for the UN/ECE METIS Group, established within the programme of work of the Conference of European Statisticians. Also available from Statistics Sweden.

Sundgren, B. (1991c) "Towards a Unified Data and Metadata System at the Australian Bureau of Statistics" Consultancy report for the Australian Bureau of Statistics (ABS). By permission of the ABS, the report is also available from the author.

Sundgren, B. (1991d) "Some Properties of Statistical Information: Pragmatics, Semantics, and Syntactics" Statistics Sweden 1991.

Sundgren, B. (1992) "Organizing the Metainformation Systems of a Statistical Office" Report for the UN/ECE METIS Group, established within the programme of work of the Conference of European Statisticians. Also available from Statistics Sweden.

R & D Reports är en för U/ADB och U/STM gemensam publikationsserie, som fr o m 1988-01-01 ersätter de tidigare "gula" och "gröna" serierna. I serien ingår även **Abstracts** (sammanfattning av metodrapporter från SCB).

R & D Reports Statistics Sweden are published by the Department of Research & Development within Statistics Sweden. Reports dealing with statistical methods have green (grön) covers. Reports dealing with EDP methods have yellow (gul) covers. In addition, abstracts are published three times a year (light brown/beige covers).

Reports published during 1992:

1992:1 (grön) Industrins konkurrenskraft och produktivitet i fokus - en utvärdering av statistiken **(Margareta Ringquist)**

1992:2 (grön) Automated Coding of Survey Responses: An International Review **(Lars Lyberg and Pat Dean)**

1992:3 (grön) TABELLER ,... TABELLER ,... TABELLER ,... - Variation och Förnyelse **(Per Nilsson)**

1992:4 (grön) Basurval vid SCB? Studier av reskostnadseffekter vid övergång till basurval **(Elisabet Berglund)**

1992:5 (beige) Abstracts I - sammanfattning av metodrapporter från SCB

1992:6 (grön) Utvärdering av framskrivningsförfarande för UVAV-statistik **(Kerstin Forssén & Bengt Rosén)**

1992:7 (grön) Cross-Classified Sampling for the Consumer Price Index **(Esbjörn Ohlsson)**

1992:8 (grön) Bortfallsbarometern nr 7 **(Mats Bergdahl, Pär Brundell, Anders Lindberg, Håkan Lindén, Peter Lundquist, Monica Rennermalm)**

1992:9 (beige) Abstracts II - sammanfattning av metodrapporter från SCB

1992:10 (gul) Organizing the Metainformation Systems of a Statistical Office **(Bo Sundgren)**

1992:11 (grön) CLAN - ett SAS-program för skattningar av medelfel **(Claes Andersson, Lennart Nordberg)**

1992:12 (grön)	KVALITETSRAPPORTEN - Utveckling av kvaliteten för SCBs statistikproduktion **(Jan Eklöf, Per Nilsson)**
1992:13 (grön)	The Use of Registers as Auxiliary Information in the Swedish Labour Force Survey **(Jan Hörngren)**
1992:14 (beige)	Abstracts III - sammanfattning av metodrapporter från SCB
1992:15 (grön)	Operationalising a Hedonic Index in an Official Price Index Program: personal computers in the Swedish import price index **(Jörgen Dalén)**
1992:16 (gul)	Some Properties of Statistical Information: Pragmatics, Semantics, and Syntactics **(Bo Sundgren)**

Kvarvarande **beige** och **gröna** exemplar av ovanstående promemorior kan rekvireras från Inga-Lill Pettersson, U/LEDN, SCB, 115 81 STOCKHOLM, eller per telefon 08-783 49 56.

Kvarvarande **gula** exemplar kan rekvireras från Ingvar Andersson, U/LEDN, SCB, 115 81 STOCKHOLM, eller per telefon 08-783 41 47.