Organizing the Metainformation Systems of a Statistical Office

Bo Sundgren

SCB

R&D Report
Statistics Sweden
Research - Methods - Development
1992:10

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 1992:10. Organizing the metainformation systems of a statistical office / Bo Sundgren. Digitaliserad av Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-1992-X101OP9210

Organizing the Metainformation Systems of a Statistical Office

Bo Sundgren

R&D Report
Statistics Sweden
Research - Methods - Development
1992:10

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

ISSN 0283-8680
Garnisonstryckeriet, Stockholm

ORGANIZING THE METAINFORMATION SYSTEMS OF A STATISTICAL OFFICE

Bo Sundgren

1992-06-30

0 Abstract

The metainformation systems of a statistical office have the role of an information infrastructure, providing a wide range of information services to persons involved in the planning, operation, use, and evaluation of surveys and other types of statistical information systems.

This report discusses some strategical and tactical efforts, which are useful to carry out, when a statistical office attempts to design and implement a metainformation infrastructure. The strategical efforts should cover (i) the development of a conceptual framework for the activities to be supported by the metainformation infrastructure; (ii) the specification of metainformation user needs; and (iii) the design of a target systems architecture for the implementation of the metainformation infrastructure.

It is not enough to have a good *strategy* for designing the metainformation systems of a statistical office. The strategy must be backed up by a well orchestrated set of tactical activities, supporting a systematical, incremental implementation of the metainformation infrastructure. The report indicates some relevant tactical activities in the following areas:

- documentation systems;
- systematical working procedures;
- computerized tools;
- metadata tapping and feeding procedures;
- policies and incentive systems.

Chapter 1 gives an overview of strategical and tactical efforts supporting the design and implementation of statistical metainformation systems. Chapters 2, 3, and 4 provide further details on each one of the three types of strategical efforts, which were mentioned above.

1 Designing statistical metainformation systems

Designing the metainformation systems of a statistical office is a very complex task. One reason for this is that the metainformation systems have the character of an **information infrastructure** of the statistical office. We shall briefly discuss the meaning and consequences of this in section 1.1.

Since it is a complex task to design statistical metainformation systems, it is of utmost importance to have a good strategy for coming to grips with it. I will propose such a strategy in section 1.2. The strategy has three components:

- developing a **conceptual framework** for the universe of interest of statistical metainformation systems, that is, developing a conceptual framework for surveys and other statistical systems;
- specifying the **user needs** to be met by a the metainformation systems of a statistical office;
- designing a **target systems architecture** for the implementation of the metainformation systems of a statistical office.

The three components of the proposed strategy will be discussed more deeply in chapters 2, 3, and 4, respectively.

However, it is not enough to have a good strategy for designing the metainformation systems of a statistical office. The strategy must be actively supported by **tactical efforts** in the form of

- a well orchestrated set of relatively small, **concrete projects and tasks**, which step by step **implement** the metainformation systems of the statistical office in accordance with **"the grand plan"** established by the strategical efforts indicated above.

If - and only if - we manage to back up the **design** of the metainformation systems with successful **implementation steps**, we are entitled to claim that we have actually **organized** the metainformation systems of the statistical office, as it is expressed in the title of this report. Notice the distinction between design, implementation, and organization.

1.1 The metainformation systems - an information infrastructure

According to Webster's dictionary an **infrastructure** is an

- *"underlying foundation or basic framework (as of a system or organization)".*

Common examples of infrastructures are roads, telephone networks, etc.

Statistical metainformation systems (like metainformation systems in general, as well as statistical information systems in general) exhibit some characteristics, which are typical for infrastructures:

- They require collective commitment and relatively large investments, which (at least initially) have to be financed by the organization as a whole.

- They have to be designed on the basis of partially unknown needs, some of which require "intelligent guesses" about the future.

- They have to be planned for a wide range of usages and users, some of whom may have conflicting needs.

- Once they exist, the marginal cost of using them is relatively low, at least in comparison with the initial investment.

Infrastructures are usually apprehended as providing a rather general **utility** of some kind. The utility often contains a rather important **service** element. Statistical metainformation systems (like metainformation systems in general, and statistical information systems in general) provide **information services**. Thus they belong to a subcategory of infrastructures that we may call **information infrastructures**.

Theories and practical experiences concerning other types of infrastructures in general, and information infrastructures in particular, are likely to provide some valuable guidance for the planning, development, implementation, operation, and usage of statistical metainformation systems as well.

1.2 Strategical efforts

As was mentioned already, the design of the statistical metainformation systems of a statistical office is a complex task, which requires a strategy. I propose a strategy, which contains three components, corresponding to three **strategical efforts**:

- *Strategical effort 1:* Develop a **conceptual framework** for the universe of interest of the metainformation infrastructure of the statistical office, that is, develop a conceptual framework for surveys and other statistical (information) systems;

- *Strategical effort 2:* Specify the **user needs** to be met by a the metainformation infrastructure of the statistical office;

- *Strategical effort 3:* Design a **target systems architecture** for the implementation of the metainformation infrastructure of the statistical office.

Each one of the components of the strategy is associated with a set of **problem solving activities** to be performed, and a set of **methods and tools** for supporting the problem solving activities.

The three sets of design problems are to some extent interdependent, so there is good reason to tackle them in parallel rather than in strict sequence. However, there is some evidence that - at least for this type of information system - it is advantageous to put emphasis on conceptual modelling in the beginning of the design process.

1.2.1 Develop a common conceptual framework

The first strategical activity to be discussed is the development of a common conceptual framework for the activities to be supported by the metainformation infrastructure of the statistical office, that is, a common conceptual framework to be used by persons involved in the planning, operation, use, and evaluation of surveys and other types of statistical systems. Such a framework could contain components like

- a common description model;
- a common "professional language";
- a formalized conceptual model.

Formalized conceptual models

Conceptual modelling has turned out to be a very useful methodology for the design of user-relevant, flexible, and viable information systems, especially for information systems with partially unknown purposes, aiming at supporting unforeseen (and to some extent unforeseeable) information needs; examples of information systems, belonging to this category, are executive information systems, decision support systems, and statistical information systems.

It seems reasonable to use conceptual modelling as a basic design methodology in the development of statistical metainformation systems. An example of what could be the result of such an approach is given in figure 1.1. The example comes from a project at the *Australian Bureau of Statistics* (ABS) [4]. See also chapter 2 of this report for a more detailed discussion.

Figure 1.1 contains an **object graph** representation of some concepts, which are fundamental in the activities to be supported by a statistical metainformation system. Since the conceptual model is associated with a *meta*information system, it could be appropriate to refer to the object graph as a **metaobject graph**. More precisely, figure 1.1 contains a proposal for a conceptual structure of a **Data Catalogue** for the ABS, the *ABS Data Catalogue* (ABSDC).

Each square box in the figure 1.1 diagram indicates a **metaobject type**, for which ABSDC should contain values of certain **metavariables**. (The metadata variables are called **metadata items** at the ABS, and they are not explicitly shown in the diagram.) The following abbreviations have been used when naming the meta-object types:

EAT Elementary Abstract Table;
POP population of objects (entities, statistical units);
SAM sample of objects from a population;
XCL crossclassification of the population into (sub)domains of interest;
PAR parameter, statistical characteristic;
DIT data item;
VAS value set;
VAL value;
SUR survey;

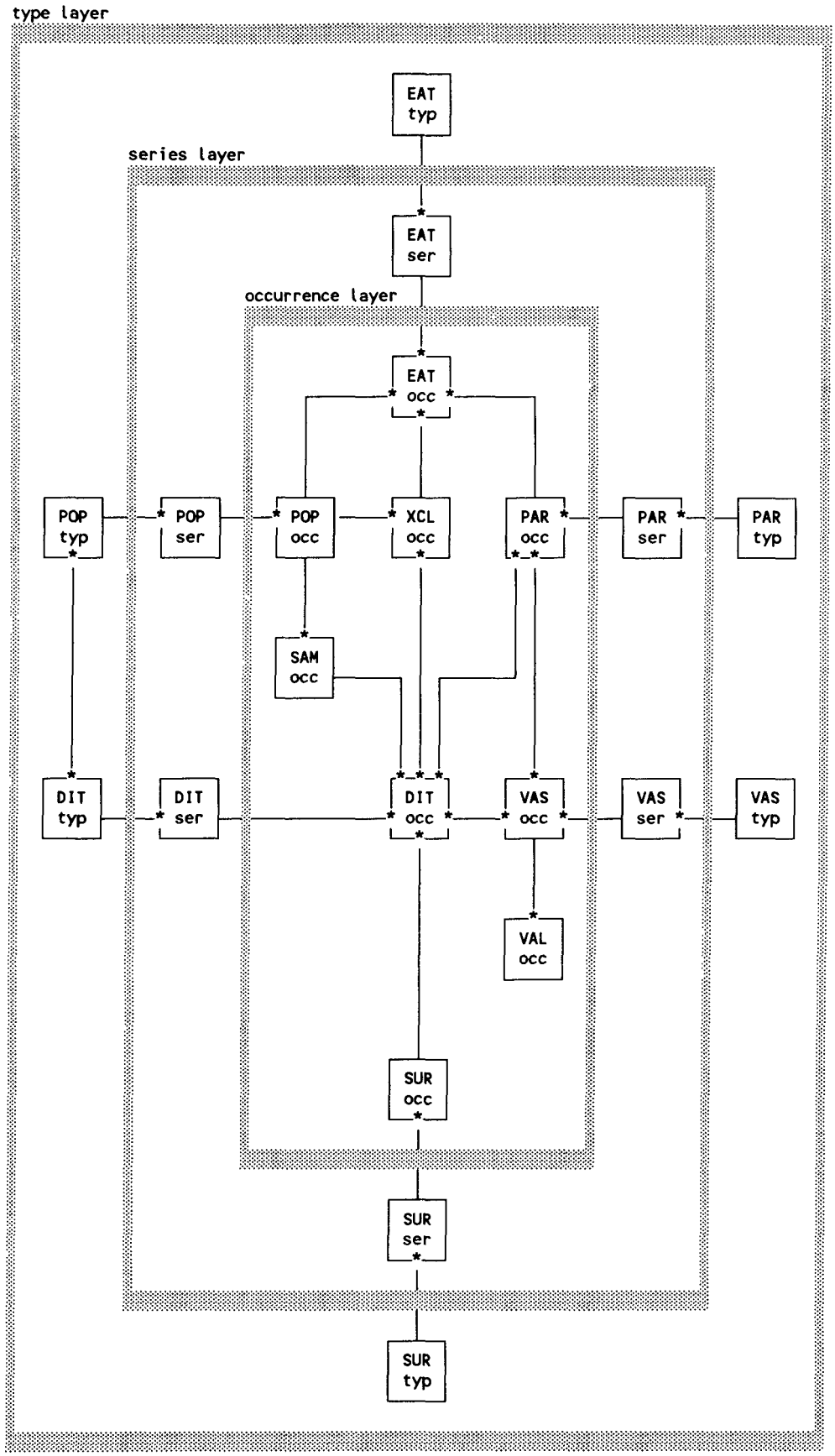


Figure 1.1. *Proposed conceptual structure of the ABS Data Catalogue. (The symbols are explained in the text.)*

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

(The terminology and notation used at the ABS differs a little from the terminology and notation used by this author in other contexts. For example, the ABS term "data item" corresponds the term "variable", as used in this report and in many other places.)

Description models and professional languages

Even for experienced information system developers, it may be felt to be a relatively abstract task to design a conceptual model for a metainformation system. Thus it may be advisable to approach the task in some pedagogical steps.

An important aspect of conceptual modelling is that it focuses on concepts (objects etc), which are central in the "business" or "activities" under consideration. In most "businesses" or "activities" these central concepts are relatively stable over time, and that is why they provide a good basis for a flexible and viable information system design.

Instead of heading directly for a formal conceptual model, like an Entity-Attribute-Relationship (EAR) model, or an ObjectPropertyRelation(time) - OPR(t) - model, one may approach the central concepts of a "business" by trying to describe the "business" verbally in some systematical way. By doing this one may find out that there is some kind of **professional language** in the organization, which contains the central concepts, with a certain degree of consensus about their definitions.

If the organization does not have a description model and/or a professional language, which is widely known and agreed upon, it may be a very constructive first step towards a formalized conceptual model to develop such a description model and language. In the context of statistical metainformation systems this means that we should try to find or develop a common language and model for describing the "business" (object system) of statistical metainformation, that is, surveys and other types of statistical systems. This approach has been tried at Statistics Sweden; a preliminary version of the result is described in [1].

We will continue our discussion of a standardized conceptual framework for the description of surveys in chapter 2.

1.2.2 Specify user needs

Conceptual modelling results in a specification of the **contents** of an information system: *"What should the information system inform about?"*. It is important for the **relevance** of the information system that it contains information about "the right things".

By specifying **user needs** as explicitly as possible, the designers of an information system will get a possibility to check once more that the planned information system will really contain user-relevant information, that is, that the **information contents** of the planned information system will correspond to important user

needs. Discussions with the users about their needs will also give possibilities to specify desirable **information system functions**, and to put **priorities** to user needs, as regards both contents aspects and functions required.

Since a metainformation system has the characteristics of an infrastructure, there are obvious difficulties to produce a complete listing of explicitly stated user needs. As was pointed out in section 1.1 there is typically a wide range of users and usages, some of which are not known, or even foreseeable, at design time. Furthermore, there may be conflicting user needs, in particular if we widen the meaning of the term "user" to include all types of "actors", who may have an interest in the functioning of the metainformation system. For example, some of the actors may mainly have to *supply* metadata, which are then *used* by other actors. In such a situation the supplier will incur costs but no benefits, whereas the user/consumer will get benefits without having to pay for them, unless some remuneration scheme is introduced.

In chapter 3 we shall introduce two methods to stimulate and systematize the process of specifying user needs. One is based upon a systematic variation of a so-called **generic query**, the other one works with **scenarios.**

1.2.3 Design a target systems architecture

It has already been pointed out several times that a statistical metainformation system is an infrastructure. Moreover, it is an infrastructure, which has to live in so-called **symbiosis** (interdependence) with other systems and infrastructures, some of which may already exist, when the metainformation infrastructure is to be designed and implemented. Consequently it is advisable to make an explicit model of the **systems architecture** of the planned metainformation system and its environment, taking into account the restrictions imposed by existing organizational and technological structures, to the extent that they cannot realistically be changed.

Even if it is possible to influence the **target environment**, where the metainformation system is to be implemented, it is usually a good idea to make an explicit model of one (or more) feasible system architecture(s), in order to test at an early stage that the proposed design is implementable in a realistical way.

Some aspects of **target architectures for statistical metainformation systems** will be further discussed in chapter 4.

1.3 Tactical efforts

As was already pointed out, it is not enough to have a good *strategy* for designing the metainformation systems of a statistical office. The strategy must be backed up by a well orchestrated set of tactical activities, which implies a systematical, incremental implementation of the metainformation infrastructure defined by the strategy.

To the author's best knowledge, there is not (yet) a statistical office in the world, which has managed to develop and implement a full-fledged metainformation infrastructure. However, there are some statistical offices, which have ambitions in this area, and which have started design and implementation activities.

Statistics Sweden is one of them, and I will use my experiences from that environment (and from some other statistical offices, which I have visited) when now going to discuss "tactical actions" in the following areas:

- documentation systems;
- systematical working procedures;
- computerized tools;
- metadata tapping and feeding procedures;
- policies and incentive systems.

1.3.1 Documentation systems

In section 1.2.1 it was pointed to the development of a common description model, a common professional language, and a common conceptual model as strategical activities in the development of a metainformation infrastructure for a statistical office. A related tactical type of activity is the development of documentation systems.

At Statistics Sweden these strategical and tactical purposes have been combined in one project. The project had the outspoken two-fold objective to

- *both* **outline a common description model of statistics production**, a model which could be accepted by subject matter statisticians, statistical methodologists, EDP specialists, and other actors involved in the design, operation, and use of surveys and other statistical systems;
- *and* **propose a documentation system for archival purposes**, that is, for describing archived microdata files from statistical surveys in such a way that they can be (re)used long after they were created and archived, even by others than those who created them.

The results of this project are presented in [1]. The proposed documentation system, which is called **SCBDOK**, is based upon the proposed description model, and there was an intensive exchange of ideas between the two lines of work during the whole project. The description model inspired the structure and contents of the documentation system, and the documentation system served as an important testing instrument for the description model.

Figure 1.2 shows the **documentation templet**, an important component of the proposed documentation system.

Both the description model and the documentation system have been applied to a number of surveys for testing purposes. The results of the tests have been satisfactory. The subject matter divisions of Statistics Sweden have requested the work to be continued, suggesting a wider scope for both the description model and the documentation system, thus aiming for

0 DOCUMENTATION STRUCTURE ETC	1 SURVEY CONTENTS
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 Archived data collections and active databases 0.4 Related documentation	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	3 COMPLETED DATA COLLECTION
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 Information sources and contact procedures 2.4.1 Data collection procedure 2.4.2 Measurement instruments 2.5 Planned observation register 2.5.1 Overview 2.5.2 OVR-matrixes with observation variables	3.1 Sampling (if applicable) 3.2 Data collection 3.2.1 Communication with the information source 3.2.2 Experiences of measurement instrument 3.2.3 *Data preparation at data collection time* 3.2.4 Non-response, causes and actions 3.2.5 Substitutions 3.3 Data preparation (coding, 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 3.4.5 Derived variables
4 STATISTICAL PROCESSING	5 DATA PROCESSING SYSTEM
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 Completed estimations 4.3.1 Point estimations 4.3.2 Estimation of sampling error (variance estimations) 4.4 Other inferences and analyses	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 and production of final observation register 5.2.1 Overview 5.2.1.1 Verbal description 5.2.1.2 System flow 5.2.2 Component descriptions 5.3 Estimations and analyses 5.3.1 Overview 5.3.1.1 Verbal description 5.3.1.2 System flow 5.3.2 Component descriptions 5.4 Result presentation and archiving 5.4.1 Overview 5.4.1.1 Verbal description 5.4.1.2 System flow 5.4.2 Component descriptions

Figure 1.2. *Documentation templet proposed as a major component of a documentation system (SCBDOK) for the microdata produced and archived by Statistics Sweden.*

- *more general statistical systems* than "traditional" surveys; and

- *more general documentation purposes* than those associated with archiving microdata for future reuse.

The documentation project work will continue in these directions.

1.3.2 Systematical working procedures

Having agreed upon a description model and a documentation system, it is a natural step to introduce systematical working procedures for the development, operation, and maintenance of surveys (and other statistical systems). 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 1.3 outlines a first version of a **systems development model**, called **SCBMOD**, which is harmonized with the proposed documentation system, SCBDOK.

1.3.3 Computerized tools

When the description model, the documentation system, and the working procedures have reached a certain degree of stability, it is appropriate and purchase, develop, or otherwize aquire software products to provide the respective activities with adequate computer support. The most basic functions should preferably be available as natural, user-friendly extensions to the word processing system and other office information system (OIS) facilities, which may be offered via the organization's local area network (LAN).

In this context, products like *Microsoft Windows* could be used to provide a uniform user interface and to integrate different software products via links.

Another possibility is to adapt commercial **CASE products** (CASE = Computer Assisted Systems Engineering), or preferably so-called **CASE shells**, adapted to the needs of statistical offices.

1.3.4 Metadata tapping and feeding procedures

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.

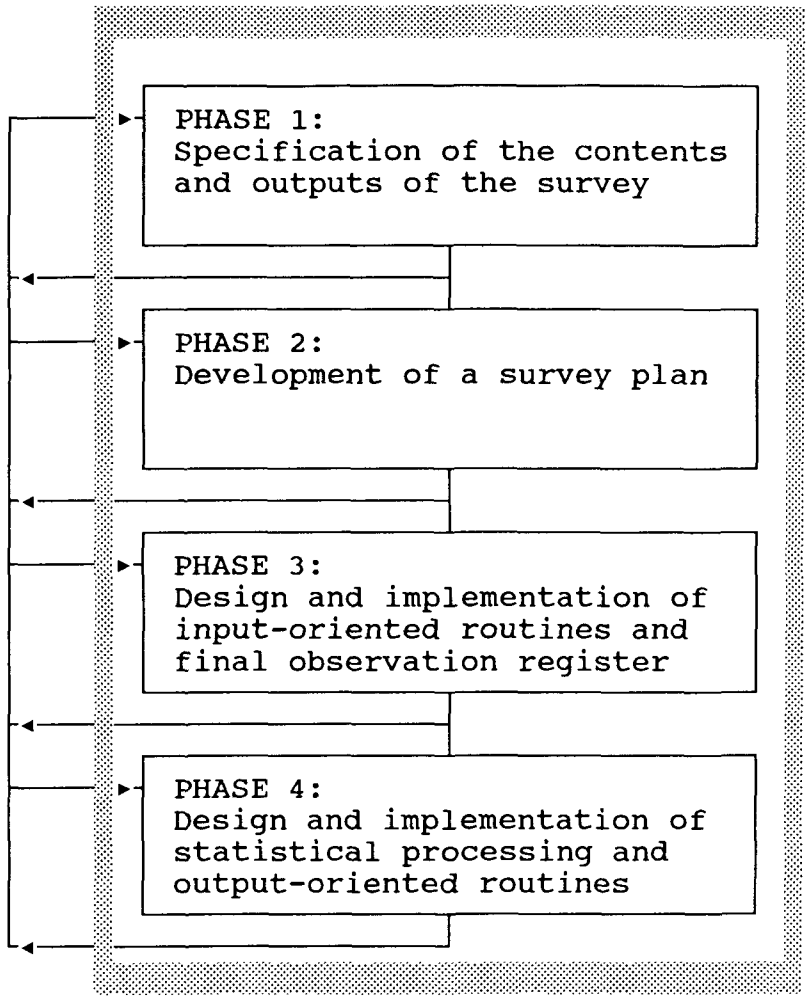


Figure 1.3. *Outline of a new systems development model for Statistics Sweden, SCBMOD.*

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, there are all the reasons in the world for statistical offices to streamline 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.**

Two examples of metadata tapping and feeding procedures are given in figures 1.4 and 1.5.

Figure 1.4 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.

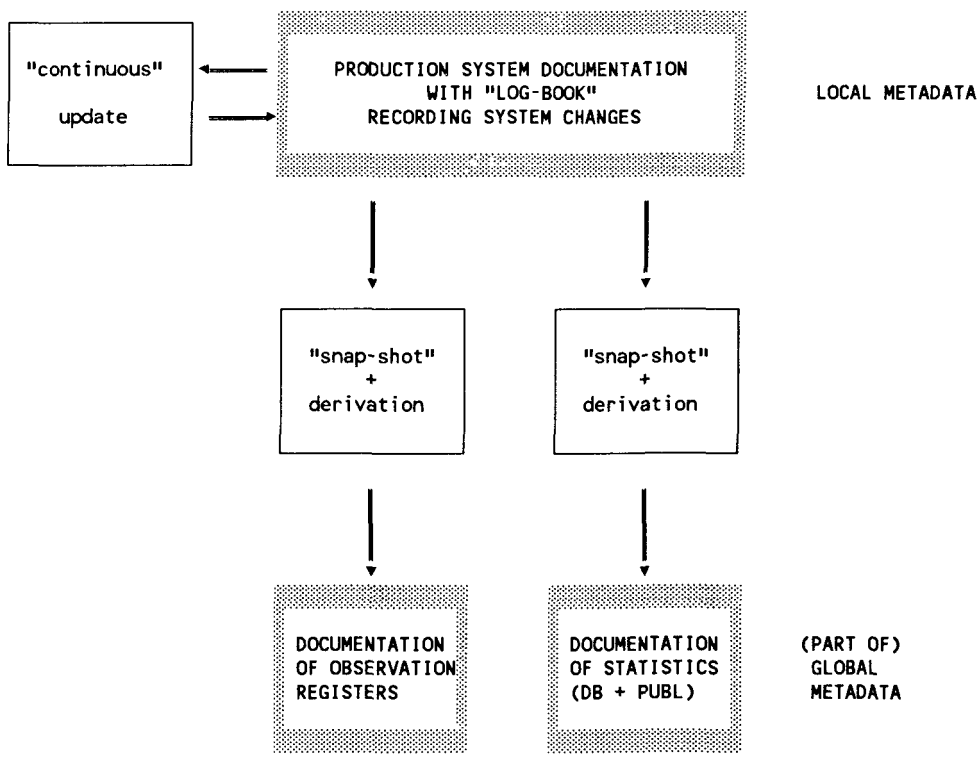


Figure 1.4. *Tapping metadata for survey end-products from production system metadata.*

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 traditional, repetitive 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**.

Figure 1.5 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.

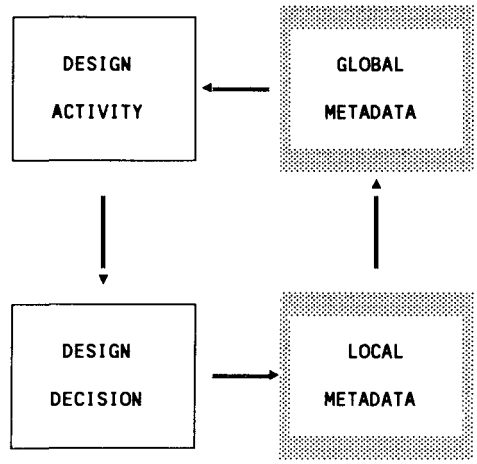


Figure 1.5. *Tapping and feeding of metadata between design activities of different surveys.*

1.3.5 Policies and incentive systems

Since the metainformation systems of a statistical office have the character of an infrastructure and a "public good", some ingenuity may have to be used to create policies and incentive systems that stimulate the development and maintenance of these systems. As in other similar situations one may choose between whips and carrots (or combinations of whips and carrots) to get the desirable stimuli.

Policies basically belong to the "whip" category of stimuli. A typical, relevant policy could be that all storage and communication of metadata should adhere to certain specified standards. Such policies could also be equipped with a "carrot" element by making sure that there are attractive software products availble, based upon the prescribed standards.

Another type of "carrot" would be to pay a certain bonus or return to those who have fullfilled certain metadata-related duties within a certain time-limit. The size of the bonus or return could be dependent upon the extent to which the metadata supplier have performed the duties with or without support from some type of corporate metadata administrator. Alternatively the bonus or return could be based upon the frequency with which the supplied metadata is used by others.

2 Developing a conceptual framework for surveys and statistical systems

Some statistical offices, for example Statistics Sweden and the Central Statistical Office of Italy, have made considerable efforts to use conceptual modelling techniques, as known from database design theory and artificial intelligence, in the development of a common conceptual framework for surveys and statistical systems. For many years these efforts were mainly focusing on data processing aspects (cf [5]), but recently the scope has been widened so as to cover aspects of statistical methodology as well; see [1], [6].

A conceptual model should primarily explain the basic concepts in the "business activities" which are (to be) supported by an information system. The "business activities" belong to the so-called **object system**, or **universe of interest**, of the information system. Since a statistical metainformation system is an information system supporting the "business activities" of a statistical office (or similar organization), the conceptual model associated with a statistical metainformation system should primarily explain the basic concepts used in the planning, operation, use, and evaluation of surveys and other types of statistical information systems. These concepts are **meta-level concepts** in relation to the **micro- and macro-level concepts**, which are explained in conceptual models associated with the statistical information systems themselves, that is, the information systems supporting the "business activities" of the users of statistics.

In connection with the design of statistical metainformation systems our main focus is on meta-level aspects of conceptual modelling. However, it is usually easier to understand meta-level concepts, if one starts from examples taken from the corresponding (object) level underlying the meta-level. In connection with statistics production this means that we should start from micro- and macro-level examples of conceptual models of object systems of statistical information systems. So this is what we shall do in sections 2.1 and 2.2 below. In section 2.3 we shall return to the meta-level aspects, and in section we shall discuss a problem complex which is particularly important for conceptual modelling of statistical systems, the problem complex of **comparability in time and space**.

2.1 Micro-level aspects

Micro-level statistical information can be formalized in terms of **e-messages** (cf [3]), that is, in terms of triples

(2.1a) $<\rho(o), \rho(p), \rho(t)>$; **"attributive e-messages";**

(2.1b) $<<\rho(o_1), ..., \rho(o_n)>, \rho(R), \rho(t)>$; **"relational e-messages";**

where

- $\rho(o)$, $\rho(o_1)$, ..., $\rho(o_n)$ are references to objects in the object system of the survey (or other type of statistical system);
- $\rho(p)$ is a reference to a property in the object system; many properties are referred to in terms of a variable and a value according to the formula: $<\rho(V) = \rho(a)>$;

- $\rho(R)$ is a reference to an n-ary object relation in the object system;

- $\rho(t)$ is a reference to a point of time or a time interval, depending on the conceptual context.

In an **OPR(t)** type of **conceptual model for a survey or statistical system** (cf [5]) the statistical information is usually categorized in terms of **e-message types**:

(2.2a) $<\rho(O), \rho(V)>$; **"attributive e-message types"**;

(2.2b) $<<\rho(O_1), ..., \rho(O_n)>, \rho(R)>$; **"relational e-message types"**;

where

- $\rho(O)$, $\rho(O_1)$, ..., $\rho(O_n)$ are references to object types;

- $\rho(V)$ is a reference to a variable;

- $\rho(R)$ is a reference to an n-ary relation.

An OPR(t) conceptual model can be visualized by means of an **object graph** (cf [5]). Figure 2.1 shows an object graph for a hypothetical survey. Figure 2.1a contains object types, object relations *and* variables of the objects belonging to the respective object types. In figure 2.1b the variables have been left out. The remaining part of the object graph gives a good overview of the *structure* of the object system of the survey, which in turn is likely to be reflected in the design of the survey itself and its database.

A conceptual model (and the object graph associated with it) is primarily a model of the so-called **object system**, or **universe of interest** of the survey or statistical system. It includes

- **objects of interest;**
- **variables of interest;**
- **object relations of interest.**

However, during the design of a survey (or statistical system) it may turn out that some objects, variables, and relations of interest are not possible to observe directly. Instead one may have to (or it may be more practical to) observe them indirectly, which means that

- *firstly*, one **observes, and collects data about**, certain **observation entities**, that is,

  - certain **observation objects;**
  - certain **observation variables;**
  - certain **observation relations;**

- *secondly*, one **derives data about** the **entities of interest** from the observation entities.

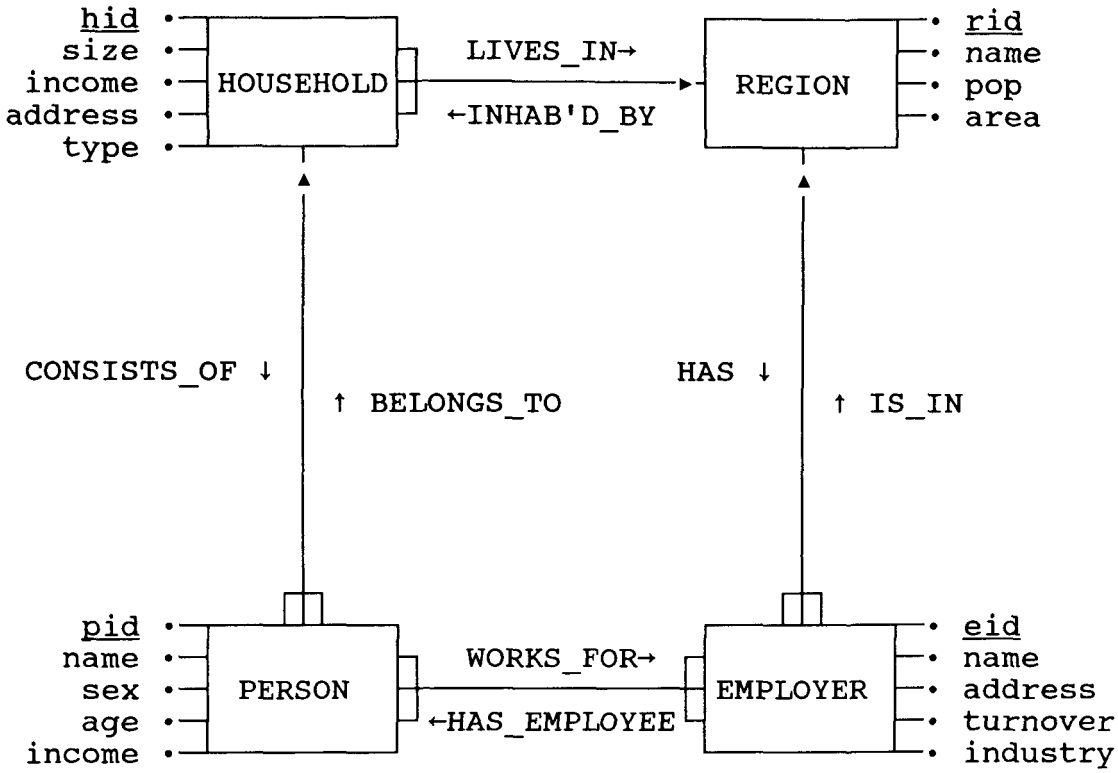


Figure 2.1a. *Object graph.*

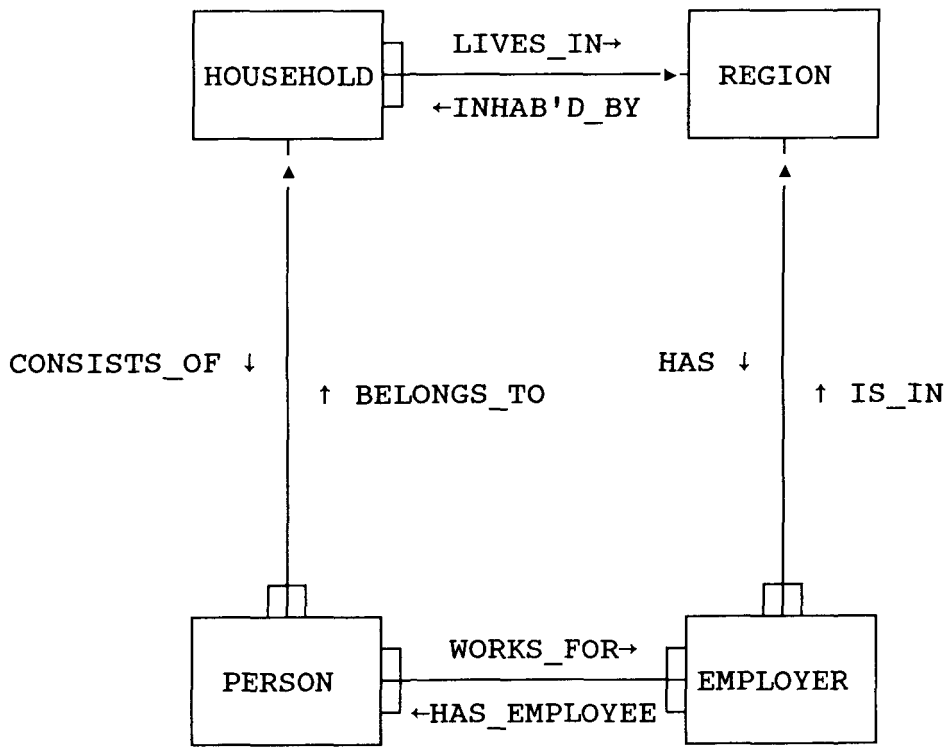


Figure 2.1b. *Object graph skeleton.*

Thus, during the design of the survey, the conceptual model, originally confined to entities of interest, may have to be **extended with** observation entities. Analogously the object graph of the survey may have to be extended.

The object graph is one way of formally representing the conceptual model of a survey (or other type of information system). The formal, **infological language INFOL,** discussed in [5], is another representation method. INFOL is particularly well suited to express in a formal and yet user-friendly way

- **formal definitions of concepts** and other types of formal relationships between the concepts of a conceptual model;

- **formal specifications of information requests,** for example so-called alfa-beta-gamma queries visavi statistical databases.

Some examples of INFOL expressions of **concept definitions** visavi the object graph in figure 2.1:

(2.3a) PERSON.address <---
BELONGS_TO.HOUSEHOLD.**address**;

(2.3b) HOUSEHOLD.size <---
CONSISTS_OF.PERSON.**count**;

(2.3c) HOUSEHOLD.income <---
CONSISTS_OF.PERSON.**sum**(income);

Some examples of INFOL expression of **information request specifications** visavi the object graph in figure 2.1:

(2.4a) PERSON(**with** income < 100000)(**by** REGION.rid * sex).
(name, HOUSEHOLD.address);

(2.4b) PERSON(**with** income < 100000)(**by** REGION.rid * sex).
(**count**, average_income <--- **sum**(income)/**count**);

2.2 Macro-level aspects

During the processing of a statistical survey, microdata are **aggregated** into macrodata. The purpose of the aggregation process is to produce **estimates** of **parameters** of some **object collectives of interest,** so-called **populations of interest** and subsets of these, which are called **domains of interest.**

The key concepts which are essential for understanding the nature of a statistical aggregation and estimation process are visualized in figure 2.2.

On the macrodata level, statistical data are data representations of **macrolevel statistical e-messages;** see [2], [3]. A macrolevel statistical e-message consists of

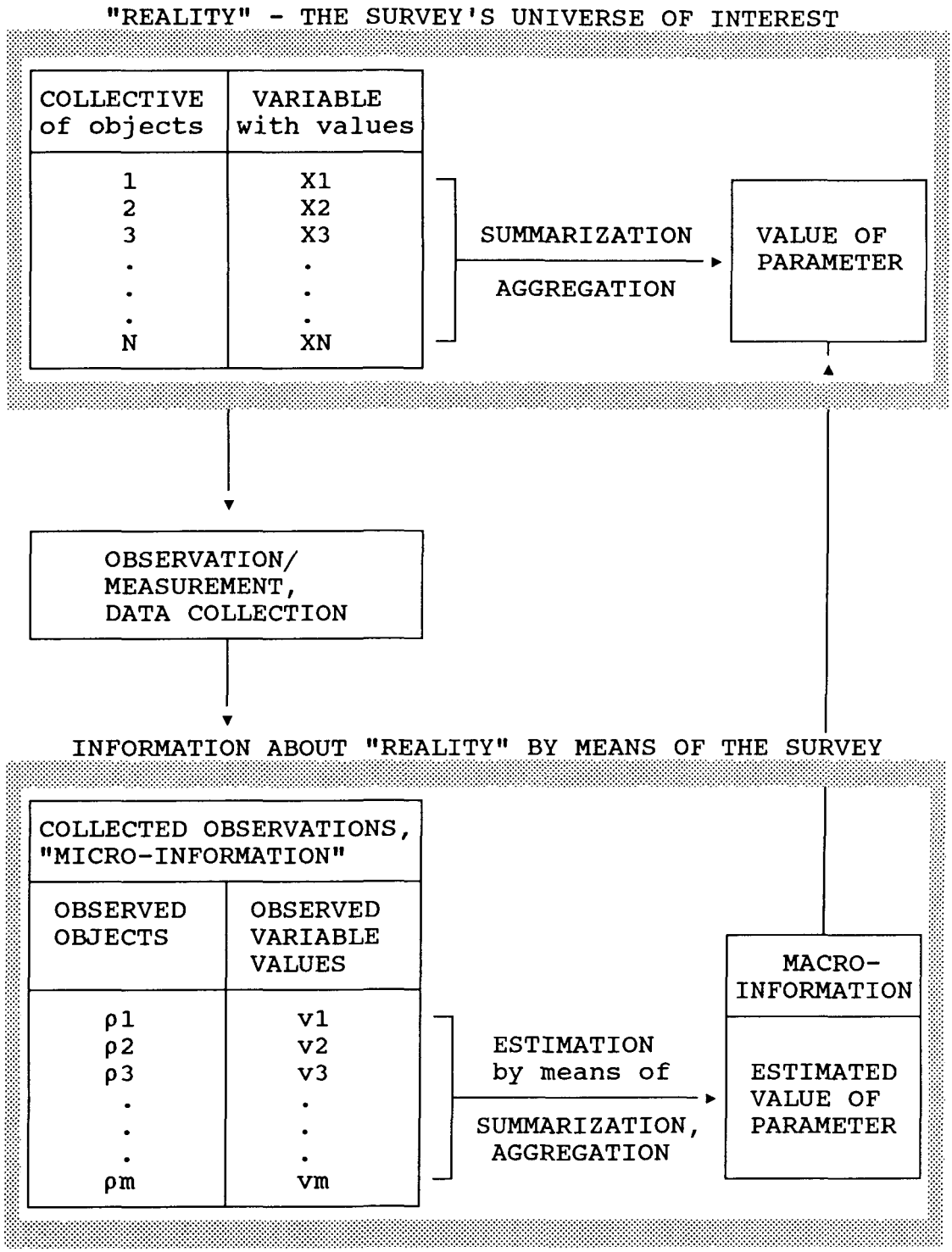


Figure 2.2. *Relationships between the universe of interest of a survey and the information about the universe of interest, which is observed, collected, and processed by means of the survey.*

- an **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 **value of a parameter**, or **statistical characteristic**, which has been 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 was supposed to have 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 macrolevel 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 aggregated statistical metainformation (macrodata) is sometimes referred to as **alfa-beta-gamma-tau analysis**; cf [2]. Figure 2.3 (covering the next three pages) shows an example of such analysis from the Australian Bureau of Statistics [4]. 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"*.

Figure 2.3. Analysis of tables in "Monthly Summary of Statistics Australia".

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 2.3 (cont'd). Analysis of tables in "Monthly Summary of Statistics Australia".

ALFA COMPONENTS	GAMMA COMPONENTS	BETA COMPONENTS	TAU COMPONENTS
Table 8 (page 2): Estimated resident population in capital cities and other major cities, 30 June 1976 to 1990 ('000).			
Resident persons in major Australian cities (...) at a certain point of time; subset of "persons"	City of residence	Count/1000	30 June 1976, 1981, 1985-1990.
Table 9a (page 3): Overseas arrivals and departures: arrivals.			
Arrivals from overseas during a certain time interval; subset of "person events"	1. Term of movement (permanent, long-term, short-term). 2. Type of arrival (settler, Australian resident, visitor).	Count	One year periods: 1987-07-01--88-06-30, 1988-07-01--89-06-30, 1989-07-01--90-06-30. Monthly: 1989-01--1991-03.
Table 9b (page 3): Overseas arrivals and departures: departures.			
Departures overseas during a certain time interval; subset of "person events"	Term and type of movement (permanent: former settlers, other residents; (long-term, short-term): (Australian residents, visitors);	Count	One year periods: 1987-07-01--88-06-30, 1988-07-01--89-06-30, 1989-07-01--90-06-30. Monthly: 1989-01--1991-03.
Table 9c (page 3): Overseas arrivals and departures: excess of arrivals over departures.			
Arrivals and departures during a certain time interval; subset of "person events"	Term of movement: (permanent and long-term, all)	Excess of arrivals over departures = sum of contribution caused by event (+1, or -1); also derivable from table 9a and 9c	One year periods: 1987-07-01--88-06-30, 1988-07-01--89-06-30, 1989-07-01--90-06-30. Monthly: 1989-01--1991-03.
Table 10 (page 4): Labour force status of the civilian population aged 15 and over, Australia.			
Civilian resident Australians at a certain point of time; subset of "persons"	1. Labour force status (standard classification) 2. Sex 3. Marital status (of females)	1. Count/1000 2. Unemployment rate (derived) 3. Participation rate (derived)	Monthly: 1990-04--1991-06.
Table 11a (page 5): Civilian labour force, seasonally adjusted series, Australia. **Table 11b (page 6): Civilian labour force, trend series, Australia.**			
Civilian resident Australians in labour force at a certain point of time; subset of "persons"	1. Labour force status (standard classification) 2. Sex 3. Marital status (of females)	1. Count/1000 (a: seasonally adjusted; b: trend estimate) 2. Unemployment rate (a: seasonally adjusted; b: trend estimate) 3. Participation rate (a: seasonally adjusted; b: trend estimate)	Monthly: 1990-04--1991-06.

Figure 2.3 (cont'd). Analysis of tables in "Monthly Summary of Statistics Australia".

ALFA COMPONENTS	GAMMA COMPONENTS	BETA COMPONENTS	TAU COMPONENTS
Table 12 (page 7): Persons receiving unemployment benefits.			
Persons receiving unemployment benefits at a certain point of time; subset of "persons"	State of residence.	Count	Yearly: June 1987-1989.
	Sex.		Monthly: 1990-04--1991-06.
Table 13 (page 7): Persons receiving sickness and special benefits, Australia.			
Persons receiving sickness benefits at a certain point of time; subset of "persons"		Count	Yearly: June 1987-1989.
Persons receiving special benefits at a certain point of time; subset of "persons"			Monthly: 1990-04--1991-06.
Table 14 (page 8): Job vacancies: industry, Australia and job vacancy rates, Australia.			
Job vacancies in Australian industry at a certain point of time;	Industry classification	1. Count/1000 2. Rate (derived; only available for the whole industry)	1990-02-16, 1990-05-18, 1990-08-17, 1990-11-16, 1991-02-15, 1991-05-17
Table 15 (page 8): Industrial disputes in progress, Australia.			
Industrial disputes in progress during a certain time period;		1. Count of all disputes in progress during the period. 2. Count of disputes which commenced during the period. 3. Thousands of workers involved in all disputes. 4. Thousands of workers involved in "new" disputes. 5. Thousands of working days lost.	Twelve months ended: 1988-12, 1989-12, 1990-12 Monthly: 1990-02--1991-04
Table 16 (page 9): Industrial disputes in progress: by industry, Australia, working days lost ('000).			
Industrial disputes in progress during a certain time period;	Industry classification.	Thousands of working days lost.	Twelve months ended: 1988-12, 1989-12, 1990-12 Monthly: 1990-02--1991-04
Table 17 (page 9): Industrial disputes in progress: states and Australia, working days lost ('000).			
Industrial disputes in progress during a certain time period.	State.	Thousands of working days lost.	Twelve months ended: 1988-12, 1989-12, 1990-12 Monthly: 1990-02--1991-04

The basic structures obtained by applying alfa-beta-gamma-tau analysis to aggregated statistical information can be referred to as **box structures** or **elementary abstract tables (EAT)**; the latter term is used in [4]:

A box structure (Elementary Abstract Table) consists of a collection of macro-level 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.

As was mentioned in the definition of the object component above, the population of interest is sometimes restricted to a subset of a larger population by means of a **selective property**, for example "female persons", "companies with more than 15 employees", etc.

2.3 Combined micro/macro-level aspects and meta-level aspects

Figure 2.2 above illustrated the key connections between micro and macro level concepts in statistical surveys and statistical information systems. Figures 2.4 and 2.5 below elaborate further on the micro/macro connections.

The **metaobject graph** in figure 2.4 is a revised version of the metaobject graph in figure 1.1, showing the proposed conceptual structure of a Data Catalogue for the Australian Bureau of Statistics. The terminology has been adapted to the terminology used in this report. Thus the **metaobject types** indicated by small squares in figure 2.4 should be interpreted as follows:

BOX "box structure" or "alfa-beta-gamma-tau structure" of macrodata;
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. (Thus an asterisk in figure 2.4 means the same as a "fork" in figure 2.1.)

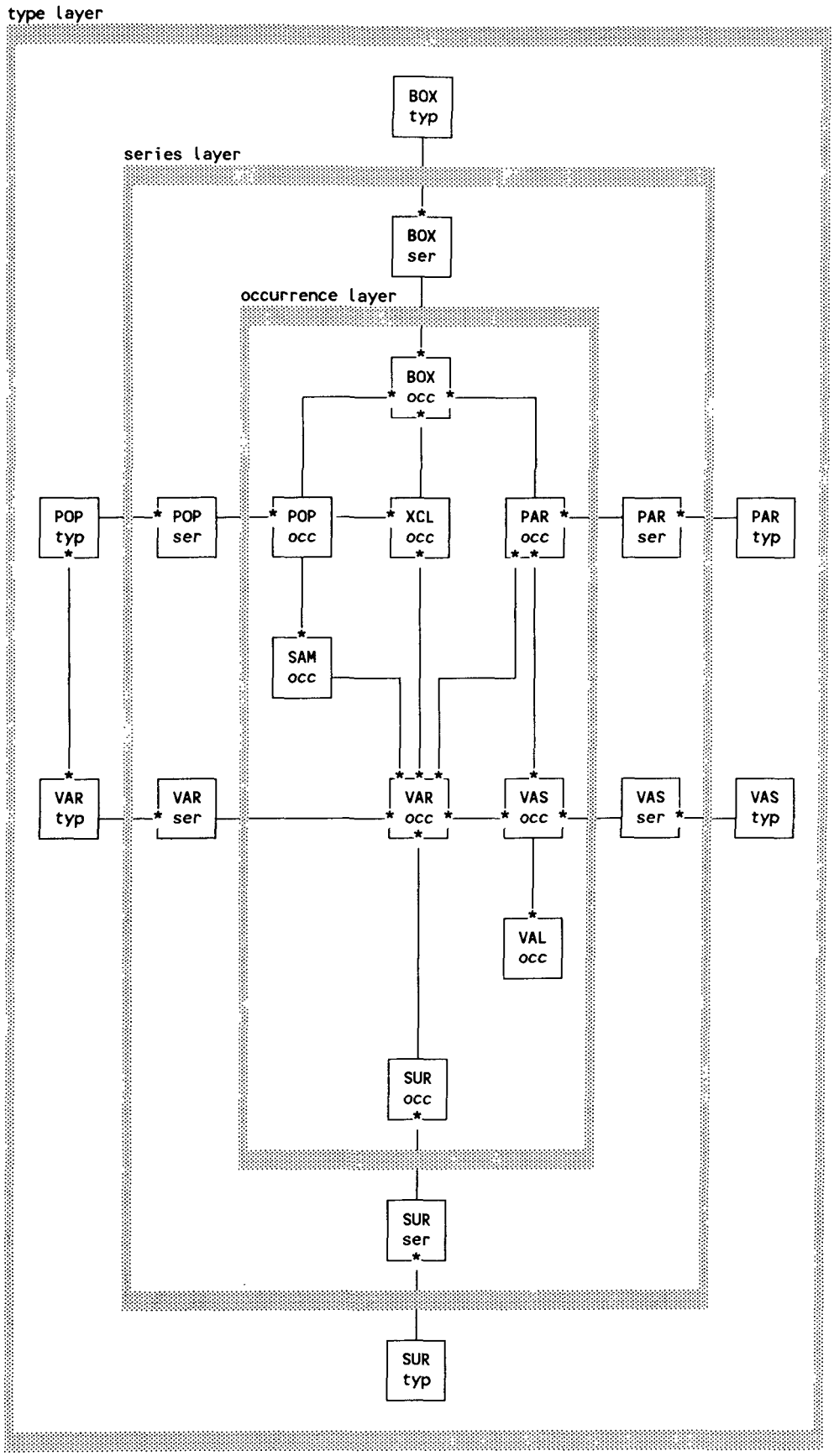


Figure 2.4. *Revised version of the metaobject graph in figure 1.1. (The symbols are explained in the text.)*

In figure 2.4 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**;

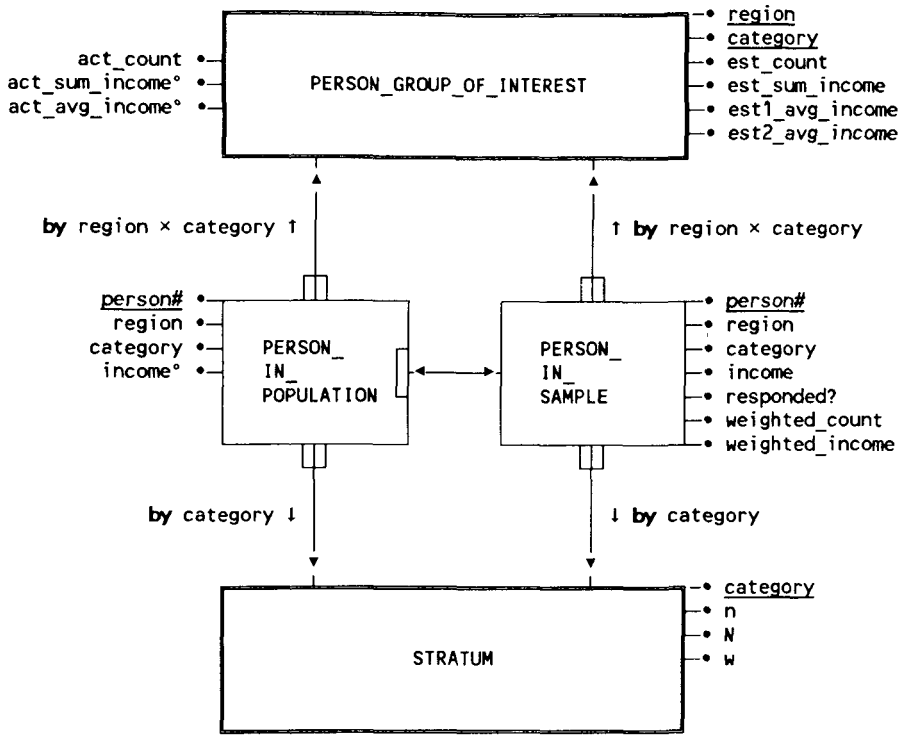
This will be further commented upon in section 2.4.

The object graph and INFOL expressions in figure 2.5 examplify and illustrate the connections between some key concepts in **sample surveys**, an important subcategory of surveys and statistical information systems.

Sampling and estimation are "twin processes" or "dual processes" in the production system of a sample survey. Figure 2.5 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 [5], [6]).

The example used in figure 2.5 is a hypothetical sample survey, where the population is a set of object instances belonging to the object type PERSON. We know the values of some variables for all the instances in the population: *person#*, *region*, and *category*. Population characteristics (parameters) that are functions of these variables can be estimated (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. Instead it should be estimated after observing a sample of PERSON objects. The sample is supposed to be 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 (generic) 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 INFOL, 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 the statistical results derived from the survey.



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 2.5. *An object graph - with accompanying INFOL definitions - corresponding to a sample 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 ring (°) after the variable name.

2.4 Comparability in time and space

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, there is no item of metadata (that is, no metadata message type) which could not be subject to change between survey repetitions. On the other hand, in practice many (maybe most) of the relevant metadata items will not change from one repetition of a survey to the next one. Both the stability and the dynamics of the metadata for a survey series must be taken into account when designing a metainformation system for a time series of similar surveys.

A failure to recognize properly the similarities as well as the dissimilarities between different survey repetitions in a survey series will negatively affect the **comparability in time**, an extremely important quality component for most 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 information usages.

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 2.4 above.

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 is similar to certain principles for **knowledge representation** used in artificial intelligence. As a matter of fact, it is also similar to the functioning of the human brain, at least according to some recent research results.

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 all metainformation, which is known to be quite different, and maybe unsystematically so, between different occurrences within the same series, or the same type, respectively. High, and relatively unsystematical variability from occurrence to occurrence within one and the same time series is typical for most **operation-based metavariables**, like "measurement problems" and "non-response rate". Most of the **design-based metavariables** will not change their values between repetitions of "the same" survey to the same extent.

Basically, the values of 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.

Why would it not be best to record all metadata (for all metavariables) on the occurrence level? Such a metadata representation would seem to be the most

"correct" and most "flexible" one. There are some counter-arguments:

(a) Occurrence level representation of relatively stable metadata implies a lot of redundance and duplication of work, in terms of (meta)data storage and (meta)data entry.

This is not one of the best counter-arguments, since data storage is relatively inexpensive, and metadata which have not changed since "last time" would not actually have to be reentered; it could just be automatically copied by means of modern wordprocessing software.

(b) Occurrence level representations do not automatically give a good overview. The user will have to process large volumes of metadata in order to get "the general picture" of things. Some of this "information value adding" processing should be done by the metadata providers and by the database administrator, preferably with the support of worksaving, computerized tools.

(c) Similarly, searches for relevant information will be very complex and resource-consuming, if all searches have to be performed on the basis of large volumes of relatively unstructured occurrence level metadata only, without being directed and supported by search methods and tools using reduced volumes of better structured metadata.

3 Finding the metainformation needs of statistical systems

In this chapter we shall discuss two methods to stimulate and systematize the process of specifying user needs. One is based upon a systematic variation of a so-called **generic query**, the other one works with **scenarios**.

3.1 Systematical variation of a generic query

A **generic query**, aiming at the specification of "all" metainformation and metadata needs of a statistical office, could be phrased along the following lines:

> "What different kinds of metainformation/metadata do different kinds of actors, in different kinds of statistical systems, need for the different kinds of activities that they perform visavi these systems."

This generic query generates a wide range of specific queries by letting four "formal variables" vary over their respective "value sets". The four formal variables are:

x1: "different kinds of metainformation/metadata";
x2: "different kinds of actors";
x3: "different kinds of statistical systems";
x4: "different kinds of activitites".

The generic query above can be represented by the quadruple

<x1, x2, x3, x4>

If we let x1, x2, x3, and x4 vary over their value sets, indepenently of each other, we shall generate a set of specific queries corresponding to the Cartesian product of the value sets of x1, x2, x3, and x4. As we shall see later, the four variables, x1, x2, x3, and x4, are not quite independent of each other, so the number of meaningful, specific queries will be slightly less than the cardinality of the Cartesian product set just mentioned. On the other hand, we shall also find out that the cardinalities of each one of the formal variables in the quadruple will be large enough to ensure that our metainformation/metadata specification process is stimulated by a multitude of meaningful specific queries.

We shall now proceed to look for possible value sets of each one of the four formal variables in the generic query.

3.1.1 Different kinds of statistical metainformation/metadata

Statistical metainformation/metadata can be categorized in several dimensions, for example, on the basis of

- whether the statistical metainformation is **product-oriented or process-oriented**, that is whether it is oriented to the products (or results) produced by statistical processes, or it is oriented to the statistical processes themselves;

- whether the statistical metainformation is **factual or rule-based**, as

discussed in [3];

- the **type of statistical metainformation/metadata** (quantitative, qualitative, free-text, etc; cf the general concept of "data types");

- the **metaobject type** (in an object graph of the statistical metainformation system), with which the metainformation is most closely associated.

There may be additional categorizations of statistical metainformation, which are meaningful and significant, but for the time being we shall confine our discussion to the four dimensions listed above.

- **Product-oriented vs process-oriented statistical metainformation**

Product-oriented metainformation focuses on the **end-products**, or **results**, of an information process, in our case the end-products or results of a statistical process or a statistical system. There are two major categories of end-products of a statistical system:

- **microdata**, often in the form of so-called **observation registers**, which are described and stored for future (re)use;

- **macrodata**, or **"statistics"**, which are described, stored and made available through databases and/or publications.

Thus product-oriented statistical metainformation will focus on either of these two types of products of a statistical production process, or on some part or component of such a product.

A product-oriented description of a statistical product could be compared with a description of a material product, which is offered to customers on a market. The potential customer of a material product will often ask for a **quality declaration** of some kind, preferably one which is delivered or approved by some respected, "objective" authority. For example, if you are considering to buy a used car, you may ask for a "status inspection report" for the car, issued by some widely respected motorist's organization. Such a report will usually give a list of **indicators**, which will give the potential customer a good overview of the quality of the product he or she is going to buy, and a possibility to judge the product's usefulness for the intended use. The prospective customer could also use the quality declaration to evaluate the price suggested by the seller.

Similarly, a product-oriented metainformation description of a microdata collection or a set of published statistics could contain a compact and comprehensive set of indicators and short descriptions of the statistical product (or some part of it). Such a description will (among other things) help a potential user of the statistical product to judge its relevance for his or her purposes.

Process-oriented metainformation focuses on how the products were actually produced. Thus process-oriented statistical metainformation will focus on how the data were originally collected, how they were prepared, processed, analyzed, and so on. Process-oriented metainformation can be said to be "HOW?"-oriented in the same sense as product-oriented metainformation can be said to be

"WHAT?"-oriented.

Process-oriented metainformation is necessary to help us remember how we actually performed certain information processes in the past, so that we can repeat the procedures, if we need to. This is a typical requirement in regular surveys, for example labour force surveys, which are periodically repeated in more or less the same way. Formally documented metainformation about work processes will also make it easier to introduce new staff into a production system, and to train the staff.

Users and reusers of published statistics and archived microdata collections will often prefer product-oriented documentation, since it gives (hopefully) most relevant and important facts about the statistical product "at a glance". It is ideally compact, well structured, and easy to understand. However, it is also common, especially for "expert users", and for users with quite new usages in mind, to request more details about how the originally collected data were actually treated during the production process. It means that they request process-oriented metainformation. There are two basic reasons behind such requests:

1. Certain things are actually both easier to explain, and easier to understand, if they are explained in a "HOW?"-oriented, rather than "WHAT?"-oriented, way.

   Example: If somebody asks for certain aspects of the quality of certain statistical data, it may be both easier and more useful for the user to describe details of editing rules and procedures than to try to give a simple indicator of the quality aspects of interest.

2. Process-oriented metainformation is in a certain sense (cf below) usually more complete than product-oriented metainformation. This completeness makes it easier to make new derivations from it, and to use it for new purposes. Process-oriented metainformation is typically "more flexible" than product-oriented metainformation.

In a previous paper [3] I discussed **infological and procedural completeness**. Product-oriented statistical metainformation (alone) could (at best) provide infological completeness, that is, it will make it possible for a user of the statistical product to make "reasonably correct" interpretations. Process-oriented statistical metainformation is essential for procedural completeness, that is, it is essential for supporting software artifacts and system administrators with the metainformation and metadata they need in order to operate and maintain the procedures of the statistical system behind the statistical products.

Process-oriented metainformation is more complete than product-oriented metainformation, in the sense that most product-oriented metainformation is derivable from process-oriented metainformation, whereas the reverse is not true. On the other hand, product-oriented metainformation has a natural appeal to users of statistics, especially to casual, non-expert users, since it is more compact and (if properly designed) relatively easy to understand and use in a "reasonably correct" way.

- **Factual vs rule-based statistical metainformation**

Factual metainformation can be formalized in terms of metalevel e-messages; cf [3] and chapter 2 of this report.

Rule-based metainformation (cf [3]) can have the form of, for example, a definition, a law, an algorithm, or a description of typical behaviour. Examples of rule-based metainformation, which may occur in a statistical metainformation system, are

- formal definitions of concepts or relationships in the object system, or in the information system, expressed in some formal language (first-order predicate logic, mathematical/statistical formulae, programming languages, etc);

- informal definitions of concepts or relationships in the object system, or in the information system, expressed in some natural or professional language;

- system flows;

- program flows and programs;

- editing and coding rules;

- work instructions for interviewers and other staff members involved in the operation of a statistical system.

- **Statistical metainformation types (cf "data types")**

The categorization into "factual" and "rule-based" metainformation provides a rough **typing** of (statistical and other) metainformation. Within each one of the major categories we may identify more precisely defined metainformation types in much the same way as data types are defined in programming languages.

Within the category of factual metainformation we may identify metainformation types like

- quantitative metainformation;
- qualitative metainformation;
- free-text factual metainformation.

Within the category of rule-based metainformation we may identify meta-information types like

- logical formulae;
- mathematical/statistical formulae;
- programming language algorithms;
- flows;
- decision tables;
- free-text rule-based metainformation (e g verbal instructions).

- **Statistical metainformation by metaobject types**

In chapter 2 we discussed how statistical information and metainformation can be specified by means of a formal, conceptual model, for example by means of an OPR(t) model with an object graph and accompanying INFOL expressions. A conceptual model offers a natural classification of metainformation kinds, namely by metaobject types. For each metaobject type one would then list the metavariables, which are of interest to observe for the metaobject instances belonging to the metaobject type.

In [4] there are tentative listings of metavariables for each one of the metaobject types in the metaobject graph, which is reproduced in figure 1.1 of this report.

3.1.2 Different kinds of actors

Different combinations of actors are involved in the different life cycle phases and activities of a statistical system. The major **actor categories** are:

- users of statistics;
- subject matter statisticians;
- statistical methodologists;
- information system methodologists;
- production systems and their operators (including software artifacts);
- managers.

The **users** are mostly located outside the statistical office. They typically liaise with subject matter statisticians and/or with the production systems and some of their operators.

The **statistical methodologists** are usually specialized on complex and mathematically oriented problems of survey design, estimation of population parameters, and statistical analysis.

The category of **information system methodologists** includes (but is not limited to) specialists on EDP technology. Information system methodologists also have special competence in structuring and analyzing complex concepts, models, and problems, a competence which is often needed in connection with statistical (and other) information systems.

The **production systems and their operators** include human, computerized, and human/computer interactive processes. The production systems of modern statistical offices are usually highly automated, and are operated via software artifacts, but they still contain some manual and/or man/computer-interactive work for handling certain routine tasks as well as exceptional situations.

The **subject matter statisticians** are supposed to act on behalf of the users. They often combine this role with parts of the roles of the other actor categories: methodologists, production system operators, and managers.

The **managers** have the responsibility to coordinate the different activities and actors within a statistical system, or within a system of such systems. In doing this they must try to make optimal trade-offs between different goals of the system,

paying special attention to those goals, which have to do with economy and timeliness. (Goals related to contents, quality, and functionality have their natural advocates in the shape of subject matter statisticians, statistical methodologists, and information system methodologists.)

3.1.3 Different kinds of statistical systems

Some experiences from Statistics Sweden indicate that essentially the same concepts and documentation templets could be used for describing the rather wide variety of surveys and statistical information systems that occur in a statistical office. Nevertheless it is often useful, not least from a pedagogical point of view, to show explicitly how a general framework for statistical meta-information can be applied to the different types of statistical systems.

Statistical systems can be categorized in several dimensions. We shall briefly discuss the following ones:

- individual surveys *vs* survey systems;
- classical surveys *vs* current registers;
- primary *vs* secondary surveys;
- input-oriented *vs* output-oriented statistical systems;
- categorizations on the basis of technology platforms.

- **Individual surveys *vs* survey systems**

In many statistical texts the term "survey" is reserved for the "classical" type of survey, which is carried out during a limited time interval, and which results in one collection of observation data. Here we shall call such a survey an **individual survey**, or a **simple survey**.

As was discussed in [3], simple surveys are often repeated periodically in more or less "the same way". Such a repetition of "similar" surveys is referred to as a **survey series**.

A survey series is one type of **survey system**, or **survey family**, that is, a colletion of "related" surveys. In a survey series the individual surveys, that is, the **survey repetitions** or the **survey rounds**, are related by the repetition mechanism. The individual surveys of a survey system may also be related on the basis of populations, topics, variables, data collection procedures, etc.

A survey system may again be a subsystem of a larger system, etc, until we reach the **survey universe** under consideration, for example the survey system of a certain statistical office, or the survey system of a country.

The concepts discussed above, and their relationships, are visualized in figure 3.1. See also [3], section 4.6, including figures 4.7 and 4.8.

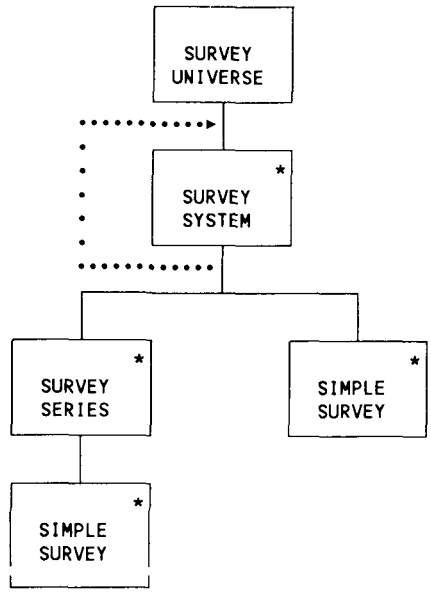


Figure 3.1. *A statistical system, or survey universe, consisting of (possibly several levels of) survey systems (survey families), where each survey system will consist of some survey series and some (non-repetitive) simple surveys. A survey series consists of simple surveys called survey repetitions or survey rounds.*

- **Classical surveys *vs* current registers**

A typical **classical survey** has the following characteristics:

- The data collection is initiated by the statistics producer, and it is done during a well-defined, limited period, which may be short, for example a week, or long, for example a year.

- After some data preparation activities, the observation data are organized in a so-called **observation register**, containing information about the observation objects of the survey.

There is another type of survey, **current registers** or **event-based systems**, having the following characteristics:

- The data collection (or reporting) is initiated by the occurrence of certain types of **events** in the object system, for example "a person migrates", "a company registers with a local tax authority", "the ownership of a car changes".

- Data are collected/reported more or less continuously, resulting in **updates** of a **current register**. It is an important aspect of the data collection to keep the current register as up-to-date as possible.

The operation of a current register type of survey is often referred to as **register maintenance**. Even if it is used for statistical purposes, a current register is often primarily maintained for administrative purposes by an administrative agency. However, statistical agencies also maintain current registers for purely statistical purposes. In particular, current registers are used by statistical agencies as **frames** in **frame procedures**. They are also used as information sources and as a basis for statistics.

- **Primary *vs* secondary surveys**

A **primary survey** is a survey, in which **primary data** are collected, and which produces statistics on the basis of these data. The statistics produced by a primary survey are called **primary statistics**.

A **secondary survey** is a survey, which is based (entirely) on statistics produced by other surveys. The statistics produced by a secondary survey are called **secondary statistics**.

Statistics from the System of National Accounts (SNA) are typical examples of secondary statistics.

- **Input-oriented *vs* output-oriented statistical systems**

Some statistical systems are **input-oriented** in the sense that they are based on one particular data collection. Other statistical systems are **output-oriented** in the sense that they focus on the particular needs of a certain group of users and aim at making available to these users all relevant statistical data, regardless of which, and how many, data collections they emanate from.

Many statistical systems try to combine input-oriented and output-oriented aspects - with varying degrees of success. A statistical office can organize the interaction between input-oriented and output-oriented aspects in different ways (cf figure 4.4 in chapter 4 of this report), but it cannot neglect any one of them, if it wants to fulfil its overall tasks in a satisfactory way. This topic will be further discussed in chapter 4.

- **Categorizations on the basis of technology platforms**

Ideally technology oriented design decisions concerning a metainformation system should (as for other information systems) be postponed until we have determined the user requirments as regards information contents and functionality of the system. Thus it may seem odd to use technological factors as a classification basis for metainformation needs. However, we have already noted that the technological artifacts themselves need metadata for their functioning, and different technological platforms will certainly have slightly different metadata needs.

Some technology-oriented choices that will affect the metadata needs are:

- systems architecture: sequential, database-oriented, ...;
- hardware platform: mainframe, mini, micro, mixed, ...;
- control: centralized, distributed, federated, ...;

- network: LAN, WAN, ...;
- operating system: ...;

3.1.4 Different kinds of activities

Different kinds of actors in different kinds of statistical systems perform different kinds of activities, and for these activities they need different kinds of meta-information/metadata support. In order to get an overview of the activities (and the metainformation/metadata needs associated with them) we may group the activities by life cycle phases and subphases (steps).

- **Life cycle phases of a statistical system**

A statistical office is involved in the life cycles of different kinds of statistical systems (surveys and survey systems, classical surveys and current registers, primary and secondary surveys, input-oriented and output-oriented systems, etc; cf section 3.1.3). When viewed on a relatively high level (cf figure 3.2), the **life cycle** of any kind of statistical system consists of three major **phases**:

- planning;
- operation and use;
- evaluation.

- **Steps of life cycle phases**

Each one of the major life cycle phases of a statistical system may be subdivided into **subphases** or **steps**. Such a subdivision will be a little different for different kinds of statistical systems. Figure 3.3 a, b, and c, illustrate a possible subdivision of the planning, operation, and evaluation phases, respectively, of a classical, simple survey.

Figure 3.4 similarly illustrates a subdivision of the operation phase of an output-oriented statistical system (cf section 3.1.3).

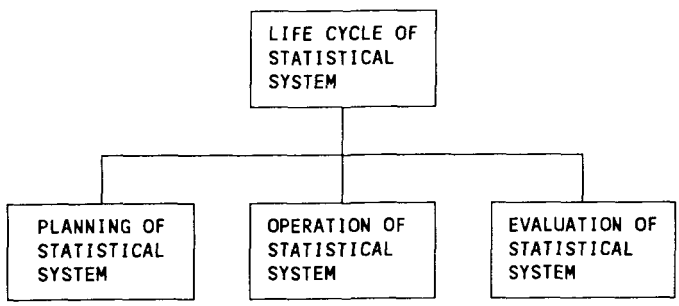


Figure 3.2. *Major life cycle phases of a statistical system.*

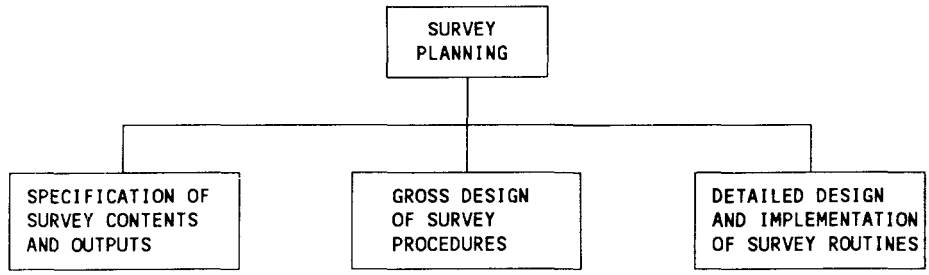


Figure 3.3a. *Survey planning steps.*

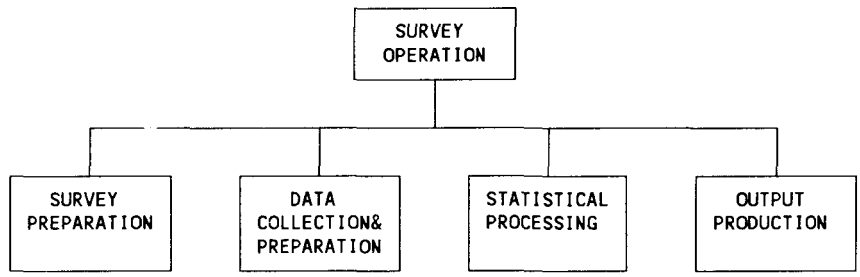


Figure 3.3b. *Survey operation steps.*

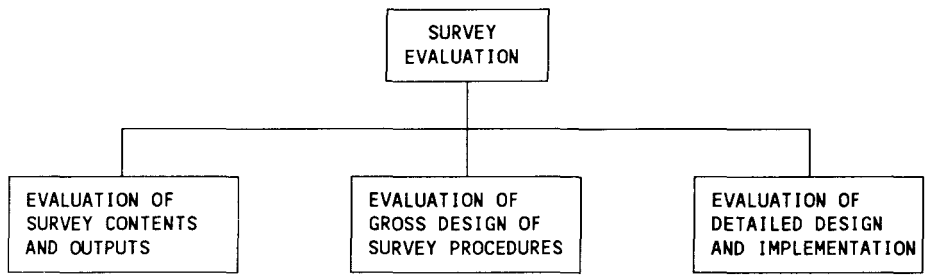


Figure 3.3c. *Survey evaluation steps.*

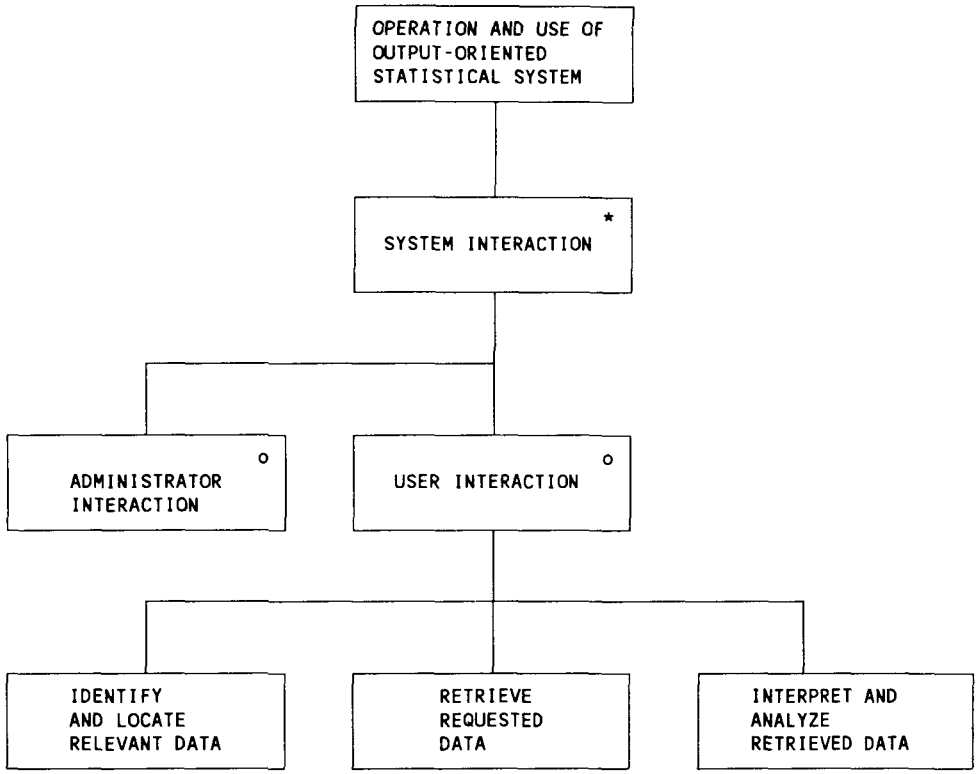


Figure 3.4. *Life-cycle diagram of the "operation and use" phase of an output-oriented statistical system.*

3.1.5 The generic query revisited

We may now summarize the contents of section 3.1 by giving a new, elaborated version of the generic query stated in the beginning of the section. The new version would read:

- *What metainformation/metadata of the following types:*

- {product-oriented, process-oriented};
- {factual, rule-based};
- {quantitative, qualitative, ...};
- {associated with metaobject type ...};

- *... do actors belonging to the following categories:*

  {users of statistics,
  subject matter statisticians,
  statistical methodologists,
  information system methodologists,
  production systems and their operators,
  managers}

- *... need visavi statistical systems of the following types:*

- {individual surveys, survey systems};
- {classical surveys, current registers};
- {primary surveys, secondary surveys};
- {input-oriented systems, output-oriented systems};
- {based upon technology platform ...};

- *... when they perform activities of the following types:*

- {phase: planning, operation and use, evaluation};
- {step: ...};

- *?*

3.2 Five scenarios of statistical metainformation usage

In section 3.1 we tried to capture metainformation needs in a relatively systematical way, by putting up a multi-dimensional space represented by a generic query with a number of variables.

In this section we shall illustrate another approach to capturing metainformation needs in a statistical office, by developing a number of scenarios of typical situations and processes, where metainformation/metadata is used.

3.2.1 Scenario 1: An end-user oriented perspective

In order to get a concrete idea of the metainformation needs of an end-user of the corporate database of a statistical office, we shall outline an end-user oriented scenario.

The scenario concerns a potential user of statistics, external or internal, who wants to find out, whether the statistical office might have some statistical data available that are relevant for the potential user's particular purpose, and if so, how to retrieve and process it, how much it would cost, how long it would take, and how to interpret the data, once it has been retrieved.

A person, who is contemplating to use statistical information from existing sources, is typically faced with four major **subtasks** (cf also figure 3.4 in the previous section of this chapter):

- to **identify** available statistical information of potential relevance;
- to **select** some statistical data for actual retrieval;
- to **retrieve** selected data;
- to **process and analyze** statistical data that have been retrieved.

Subtask 1: Identify available statistical information of potential relevance

We assume that the user has a particular task to perform, and that the user has a general idea that statistical information may be instrumental in performing this task. The task may be

- to solve a specific problem of some kind;

- to collect, analyze, and present data as a basis for a forthcoming decision to be taken (or an already implemented decision to be evaluated) by a political body, a company, or some other type of organization;

- to carry out research on a particular topic;

or the like.

The first question from the potential user of statistical information will be read something like:

(a) What statistical data may at all be available, statistical data that are relevant for my (the user's) task?

In order to identify available statistical information of potential relevance, the user needs to have (or to get)

- a reasonably well conceptualized idea of his/her own problem or task;
- some ideas concerning what statistical information might be helpful;
- a directory to available statistical information;
- tools for searching and navigating in the directory.

This points to the need for

- a **data catalogue**, or **directory**, giving a good overview of the whole information potential of the statistical office, including

  - what is **directly** available from the corporate database of the statistical office itself, through its **macrodata** and **microdata** components;

    (this role of the data catalogue is sometimes referred to as the **active role**);

  - what more is possibly available, **indirectly**, through person-to-person contacts with the staff of individual surveys;

    (this role of the data catalogue is sometimes referred to as the **passive role**);

- different **search mechanisms** that enable the user to search actively for information concerning a particular topic.

As for **data catalogues**, from the user's point of view there should ideally be *one* directory covering "all" statistical information of potential relevance to the user's problem, even if the statistical data themselves are stored in many different places, and/or "are owned by" (are the responsibility of) several organizational units within (or even outside) the statistical office under consideration. Thus there may be a so-called **"structure clash"** between the statistics user's and the statistics producer's perspective, since the producer's perspective is usually limited to a relatively small number of "input-related" surveys.

The **search mechanisms** need to be of several different types. The "traditional" way of searching and navigating in statistical databases and their directories has been through **hierarchically organized menu systems**. This type of search mechanism may be satisfactory for many users, especially for **casual users** with rather "standard" information needs.

However, searches based upon hierarchically organized menu systems may sometimes be felt to be rather rigid and inefficient. On the one hand, there are the users who know rather precisely what they are looking for, and they find it rather boring to go through the hierarchies; they may prefer to state their requirements through a command language, or by simply giving an identification of, say, a particular time series that they are interested in.

On the other hand, there are the users who are not very articulated about what

they are looking for, and they may find a menu hierarchy too rigid, imposing a view of the statistical information that they do not feel "at home" with. The user may have a view of the world, which is structurally different from the particular hierarchical view imposed by a hierarchical search mechanism. Such a user may prefer to view the world through an alternative hierarchical model, or through a model which is not hierarchical at all. Moreover, *one* hierarchy may not be enough to direct the user to all relevant statistics. Consider for example a user who is interested in the value of the investments that have been made in hospital buildings. (Should he/she navigate via "building statistics" and/or via "health statistics"?)

More flexible search mechanisms may be based on **key-word searches** and other forms of **free-text searches**, supported by a good **statistical thesaurus**. Statistical thesauri are currently being developed by the Australian Bureau of Statistics (Suzanne Ridley) as well as by Statistics Sweden (Malkon Lindmark).

Several interesting **research projects** have been carried out concerning how to organize well-structured overviews and flexible search mechanisms for statistical databases. Some of the results from these projects have been reported at international conferences, in particular at the conferences on **Statistical and Scientific Data Base Management (SSDBM)**, which have been held for about a decade now.

The **text mass**, upon which structured and less structured searches for relevant statistical information may be carried out, will consist of several parts. It should be noted that a free-text search may have to cover several metavariables in the data catalogue. For example, if a user is interested in the production of refrigerators, it may not be sufficient to search a metavariable like "table title", since the word "refrigerator" may rather be found as the textual name of a value in a value set or classification. Thus the search should cover (at least) both "table titles" and "value set value names".

Another part of the text mass to be (sometimes) utilized by search mechanisms will be stored in, or derived from, the local **survey knowledge bases**. Major parts of this text mass will probably be less formalized than the text mass contained in the corporate data catalogue. Naturally a search scanning through the whole text mass of all survey knowledge bases would be very resource-consuming, so more intelligent search strategies will have to be developed.

There seems to be virtually no limit to how sophisticated and how tailored to the needs of different users that data catalogues and search mechanisms may ultimately be. However, it is essential to plan the development of the corporate database of a statistical office in such a way that the development of data catalogues and search mechanisms can be done **incrementally** (step by step) and as **independently** as is logically possible from the development of the database itself.

The development of a suite of search mechanisms for a statistical office could start with rather simple tools similar to those, which have already been in use for some time. New contributions can come from developments of new output-oriented systems, provided that all new developments adhere to **common standards** and are designed to have such a **high degree of generality** as to be

useful also for other projects, focusing on other users and other parts of the information potential.

Two conclusions:

- The directory, or **"metainformation gateway"**, to the statistical database(s) and files of a statistical office should ideally cover "all" statistical data of potential relevance to the category of users, for which the gateway is designed, including even some statistical data produced by other organizations than the statistical office under consideration.

- There should be several different types of tools and mechanism for searching and navigating in the directory, including some rather sophisticated associative, thesaurus-supported, highly interactive ones, where the imaginativeness of both the user and the producer can be constructively exploited.

Subtask 2: *Evaluating the usefulness, costs, and availability of the potentially relevant statistical information identified in subtask 1*

Let us now assume that the user has identified data, which seem to be relevant for his or her task. The user would then like to investigate such things as

(b) Is the quality of available data sufficient for my purposes?

(c) Are data emanating from different sources, and different time periods, comparable with one another?

(d) How can relevant data be retrieved and processed? Is the data catalogue an **active** one, that is, can I just "press the button" once I have decided what I am interested in, or is it **passive**, so that I will have to make personal contacts before being able to retrieve the data? Can I download data and metadata to my own systems for further processing?

(e) Are there any secrecy constraints affecting the retrieval and processing of relevant data, and, if so, can they be overcome?

(f) How much would it cost to retrieve and process the relevant data, and how long it would take to satisfy the request? Will the benefits of getting the data justify the time and costs of getting them? What less expensive, and/or less time-consuming alternatives may be available?

It should be possible to give at least rough, approximate answers to most of these question by consulting the corporate **data catalogue**, with its relatively well structured metadata, which in an ideal future system will have been more or less automatically extracted, or "filtered" from the survey knowledge bases.

However, sometimes the user will find that the answers are not detailed or precise enough. Then he or she should be able to "dig into" the **survey knowledge bases** themselves, that is, the survey knowledge bases of the surveys underlying the statistical information of potential interest to the user.

For example, in order to judge the quality of certain statistical data the user may need to know precisely how a certain survey was designed in certain respects, or more details about certain problems that were encountered during the meaurement process. There may be evaluation studies of the surveys that may be of interest for a sophisticated user. Etc.

Subtask 3: Submitting a request

When the user has decided upon a certain request for data, he or she would like to have the request processed:

(g) Process my request, as I have now defined it (or: as I will now define it).

In most cases it should be possible to process the request directly, on the basis of the contents of the corporate database of the statistical office, but sometimes the user may have to be referred to person-to-person contacts.

In particular, person-to-person contacts may be necessary,

- if certain **confidentiality problems** are present; or
- if certain **infrequently requested data** have to be "activated".

Subtask 4: Analyzing the results of a request

When the results of a request have been presented to the user, he or she may need additional help to interpret the results, and/or to formulate new requests:

(h) How should I interpret ...? What is the definition of ...? How can I further analyze ...? Please, perform the following analysis ...

Once again many of these requests for additonal help should be answerable on the basis of the contents of the corporate data catalogue, but once again it may also sometimes be necessary to refer the user back to local survey knowledge bases.

When going from "simple" retrieval operations to more or less complicated **statistical analyses**, we encounter another interesting potential component of future systems. For each statistical method, and each software tool, there may potentially be a **statistical expert system** assisting naive and/or experienced users of statistical data.

Some statistical offices have already successfully developed various expert systems, or **knowledge-based tools**, as they are often (and more modestly) called today. In the future such tools should be integrated with the data/metadata infrastructure of statistical offices. Well-defined **interfaces** between expert systems and the data/metadata systems have to be established.

3.2.2 Scenario 2: A production-oriented perspective

For quite natural reasons a statistics producer's metainformation needs differ quite significantly from those of a statistics user, as regards both scope and contents. As regards the scope the difference is so significant that it is right to

talk about a "structure clash", as we did earlier. The metainformation interests of a particular, statistics-producing organizational unit is usually limited to one survey or survey series, or possibly to a small number of input-related surveys for which the organizational unit has the responsibility. Many statistical offices are organized in such a way that the statistics-producing units also have the responsibility for (some) user-oriented aspects of the statistics produced. Nevertheless, it is neither common, nor easy, for a basically production-oriented unit to take responsibility for the needs of different users to relate the statistics produced by this unit to statistics produced by other units, or even by other organizations.

The metainformation needs of a statistics producing unit emanate from the unit's responsibility to perform tasks like

- operating certain surveys in an efficient way;
- maintaining the survey production systems, including EDP systems;
- introducing and training new staff in the work routines of the unit.

A modern survey production system is operated by manual and computerized routines, often in quite close and sophisticated interaction. We may classify the routines as

- **manual,**
- **automatical**, or
- **interactive,**

depending on the role of human beings and computers in the respective routines. All three types of routines require metainformation and/or metadata. For manual and automatical routines, the requirements have become relatively well defined over the years, but for interactive routines, being a more recent phenomenon, the needs for formalized metainformation/metadata are not (yet) equally well established.

Metadata to be used by computers still have to be relatively formalized, usually according to some format or syntax. Metadata are sometimes said to **"drive"** computerized routines. Analogously one could possibly say that there are certain types of metadata, which are used for "driving" the manual and interactive routines of a statistical production system.

There is no sharp borderline between maintaining and (re)designing a statistical production system. It is customary to refer to a small number of minor changes in the routines and data holdings of a production system as "maintenance", whereas the term "redesign" is used, when a relatively large number of changes, and/or relatively significant changes, are made. In principle, the metainformation/metadata needed is of the same nature, regardless of whether the changes in the production system are called maintenance or (re)design; thus we refer to the discussion in the next section.

3.2.3 Scenario 3: A design-oriented perspective

The design of a statistical system (survey or statistical information system) can be organized in a number of phases and steps. The phases and steps do not necessarily have to be carried out in strict sequence. Results achieved in later

design steps may sometimes necessiate revisions of (preliminary) design decisions taken earlier in the design process, so that the process becomes iterative rather than sequential. A number of design issues may also be so interrelated that some kind of "simultaneous optimization" must be aimed at; example: the design steps concerning sampling and estimation procedures. Nevertheless, it seems useful for a statistical office to establish some kind of standardized design methodology, including a "checklist" of phases and steps that have to be considered.

A checklist for the design of statistical systems could include phases like (cf figure 1.3 in chapter 1):

Phase 1 Overview of the contents of the statistical system in terms of

- object system of interest;
- major statistical information outputs.

Phase 2 Overview of the production system in terms of

- survey plan(s);
- processing plan(s);
- presentation and distribution plan(s).

Phase 3 Detailed design and implementation of input-oriented subsystems and microdata holdings.

Phase 4 Detailed design and implementation of subsystems and tools for aggregation, estimation, statistical analysis, and output-oriented processing, storage, and presentation.

Through all its phases and steps the design process is both a user and producer of metadata. Thus the collective of design processes is in a sense the key to an efficient, integrated metainformation system of a statistical office.

As an illustration, consider a team of staff members, who are about to design a new survey, or to modify an existing design. The team consists of a statistician, a subject matter expert, and an EDP specialist, and the design should cover aspects of contents, statistical methodology, and EDP.

In many cases the survey to be designed will be a new version of a survey that has already been carried out. The existing specification of the previous version of the survey will then be an obvious starting point for the new (modified) design.

Even in situations, where the survey to be designed is a new one in a more genuine sense, there will probably be "similar" surveys, from which design experiences can be used.

Thus the first question would be:

(a) Which is (are) the most "similar" survey(s), which we can take as a starting point for our (the design team's) design work?

Then the natural request would be:

(b) Retrieve relevant design documents for the "similar" survey(s), and place them in the active database of our design environment; for example: in the "working database" of our **statistical CASE tool.**

(CASE = Computer-Aided Systems Engineering; "statistical CASE" = Computer-Aided Survey Engineering.)

Our design team will have to tackle a range of design problems, implying tasks like

(c) Prepare design proposals concerning

(i) conceptual model (objects and variables of interest, desirable outputs);
(ii) frame and sampling procedures;
(iii) questionnaire and other measurment instruments and data sources;
(iv) coding rules and classifications to be used;
(v) non-response handling, data editing rules and procedures, observation modelling;
(vi) aggregation and estimation procedures;
(vii) presentation, distribution, archiving;
(viii) data processing system.

Different (combinations of) team members will work on different tasks, to a large extent in parallel and iteratively. Proposals in one design area will often affect those in another one.

Some design proposals may have to be examined for coordination and policy reasons. Ideally the statistical CASE tool should

(d) Call the design team's attention to proposed design decisions that may need policy considerations, or even formalized approval.

3.2.4 Scenario 4: A managerial perspective

A car manufacturing company produces cars. A statistical office produces statistics, that is, a certain type of information. Both the managers of a car manufacturing company and the managers of a statistical office need information about their respective production processes in order control, coordinate, and evaluate the business activities. Since the production processes of a statistical office are information processes, the information systems needed by managers of a statistical office will typically be metainformation systems. When organizing its metainformation systems, a statistical office should therefore consider the (meta)information needs of its managers on different levels, especially if the statistical office is aiming at an integrated system, or infrastructure, of meta-information systems, where redundant duplications of metadata and metadata handling operations are to be avoided.

There are several management levels of a statistical office. On the lowest managerial level, there is typically a person who is responsible for a single survey (series), or a small set of related surveys. On intermediate levels there are typically managers who are responsible for

- a set of input-related surveys; and/or
- a set of output-related user categories.

On the highest management level the rationality and efficiency of the whole organization must be taken into account. Among other things the top management of a statistical office must

- establish budgets and follow up the performance of different parts of the organization in economical terms;

- establish quality standards and follow up the quality of different statistical products;

- coordinate the statistical products and the production processes in different parts of the organization (for example, the collection-oriented and the user-oriented parts);

- evaluate user satisfaction with the products and services of the statistical office.

In order to perform these tasks the mangers will require the metainformation systems to produce and make available

- economical information, including productivity data;

- quality information, covering several quality dimensions;

- an overview of the contents of and relations between different statistical products and different production systems, including an overview of concepts, definitions, and standards;

- statistical information about statistics requested and retrieved (if available), indicating among other things

    . statistics produced, which have not been requested;
    . statistics requested, which have not been available;

- user evalutations of statistical products.

3.2.5 Scenario 5: A technical perspective

When discussing the production-oriented perspective (scenario 2) we noted that the computer-supported processes of a statistics production system have metadata needs, which have to be taken into consideration when designing statistical metainformation systems. Similarly the metainformation systems themselves have metadata needs, or "metameta"data needs.

When applying a technical perspective to the specification of metainformation needs of a statistical office, we systematically consider and analyze the metadata needs of the technical subsystems of the data/metadata infrastructure, for example the metadata needs of different software components. We shall not pursue the technical perspective further here.

4 Specifying a target architecture for the metadata infrastructure

When designing an information system it is usually more practical to start the design process from some kind of **target architecture** than to start completely from scratch. The target architecture can be seen as a collection of **constraints**, which should be adhered to during the design process. There may be different kinds of reasons behind different constraints, for example:

- **Experiences from designing similar systems** in the past may suggest that it is wize to follow certain rules and to use certain components, subsystems, and subsystem structures;

- The information system to be designed may have to perform **interactions with other information systems**, which are already in operation, and which have certain given interfaces to their environment;

- The information system to be designed will have to be implemented by means of an **existing technical infrastructure**, which in practice cannot be affected (for the next few years).

Reasons like these will certainly exist in most situations, where the meta-information systems of a statistical office are to be designed. Thus one of the first actions in such a situation should be to summarize the constraints in the form of a target architecture.

In this chapter we shall discuss certain typical features of a target architecture for statistical metainformation systems. The features are grouped in four major categories:

- integration of data management and metadata management;

- database orientation and sharing of data/metadata within and between systems;

- data/metadata interactions between local, intermediate, and global levels;

- data/metadata exchange between input-oriented and output-oriented systems.

Having discussed each one of these feature categories separately, we shall outline a more comprehensive target architecture, which contains a combination of the features introduced. Finally we shall discuss the needs of harmonizing the metainformation infrastructure with other infrastructures of the statistical office, for example its Office Information System (OIS), and how such a harmonization could affect the target architecture.

4.1 Integration of data management and metadata management

The first important thing to note when designing the architecture of a metadata infrastructure is that it must be coordinated with the architecture of the corresponding data infrastructure. In a statistical office, every activity, which somehow manages data, should also manage the metadata, which is naturally associated

with the data; cf figure 4.1, which is a revised version of a figure in [3].

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, should compose readable tables, he or she would have to (re)introduce metadata, explaining the meaning of the (object) 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.

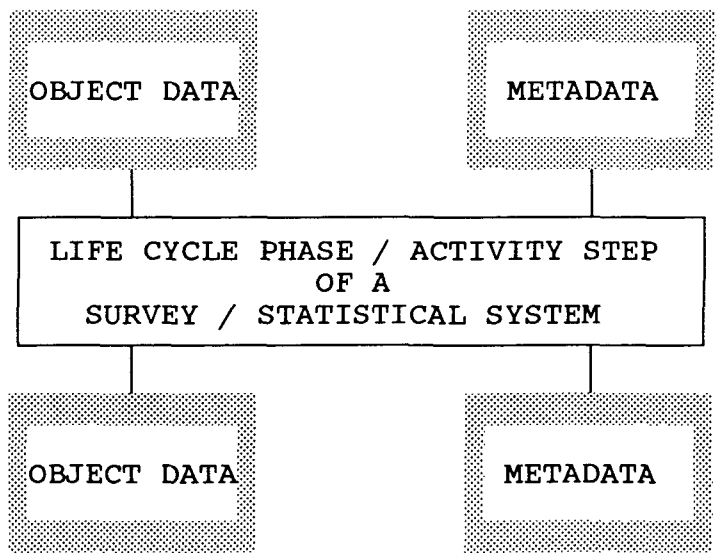


Figure 4.1. *Every life cycle phase and activity step of a statistical system should be designed so as to not only use, but also produce, metadata associated with the object data used and produced.*

Figure 4.1 stresses the fact that every life cycle phase and activity step of any survey, and any other type of statistical system, should be designed so as to use, process, and produce metadata in parallel with, and integrated with, its usage, processing, and production of (statistical) (object) data.

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

It should be a challenge to every statistical office to organize its metainformation production, storage, and use 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.

4.2 Sharing of data/metadata within and between systems

It follows from the observations in the previous section that **sharing of metadata** (and 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.

The sharing of metadata (and object data) between different surveys, other statistical systems, and non-statistical systems is made easier, and easier to give adequate computer-support, if the **gross architecture** of the data/metadata management of a statistical office, as well as the architecture of the individual production systems, and auxiliary systems, is designed to be **database oriented.**

Figure 4.2 illustrates the difference between a traditional, sequential systems architecture and a modern architecture based upon database oriented principles, by applying the two architectural schemes to the productions system of a "classical" statistical survey.

The traditional, sequential architecture of a survey production system is visualized by figure 4.2a. This type of architecture was in fact the only possible one, as long as the technology of statistical offices was characterized by serial storage media (magnetic tapes) and batch processing.

Figure 4.2a does not explicitly recognize the metadata handling in the survey processing. Neither does it indicate any links between the survey under consideration and other surveys and information systems. Figure 4.2b has been improved in these respects.

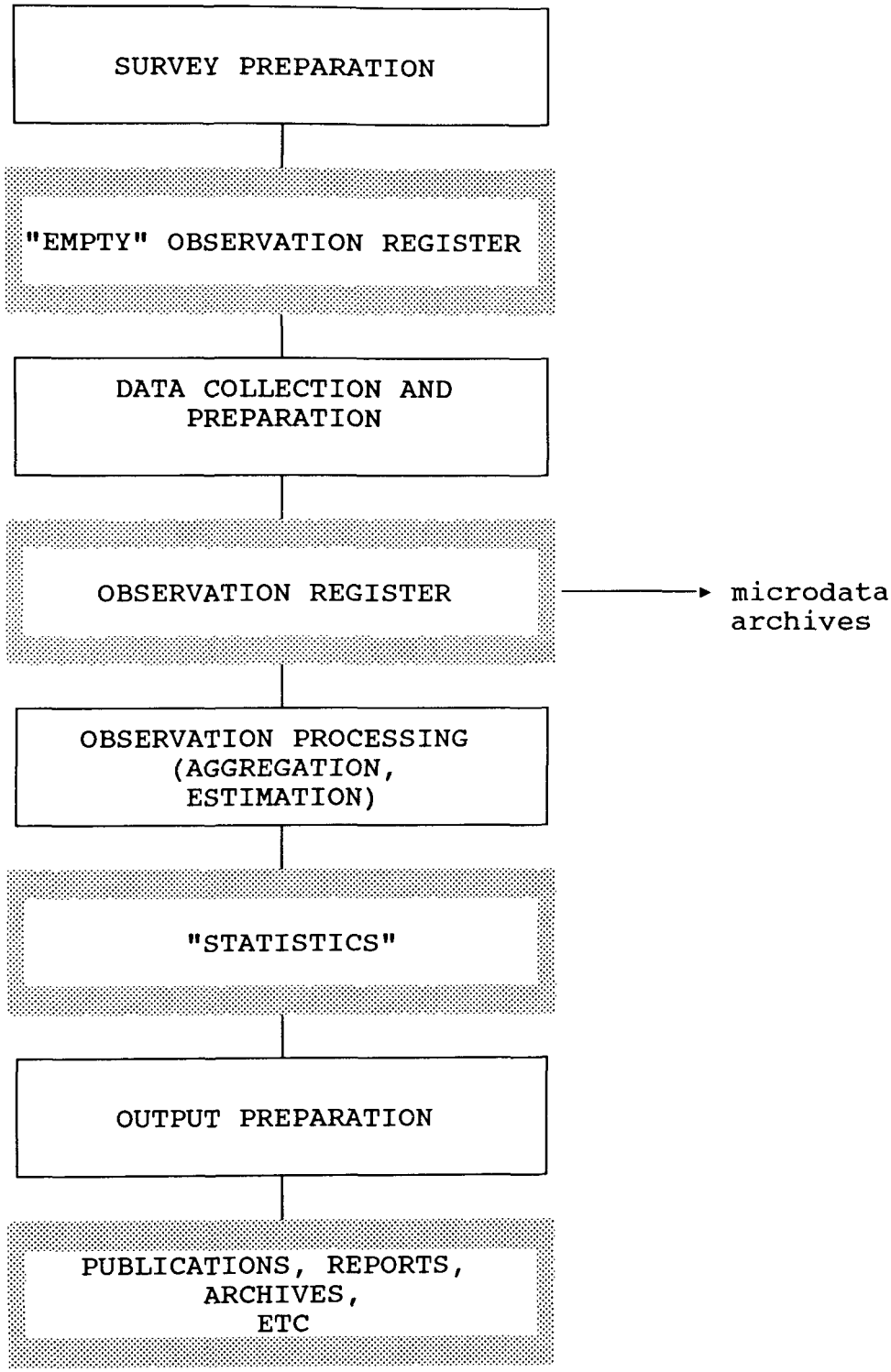


Figure 4.2a. *Survey production system: sequential architecture without explicit recognition of metadata processing and links to other surveys.*

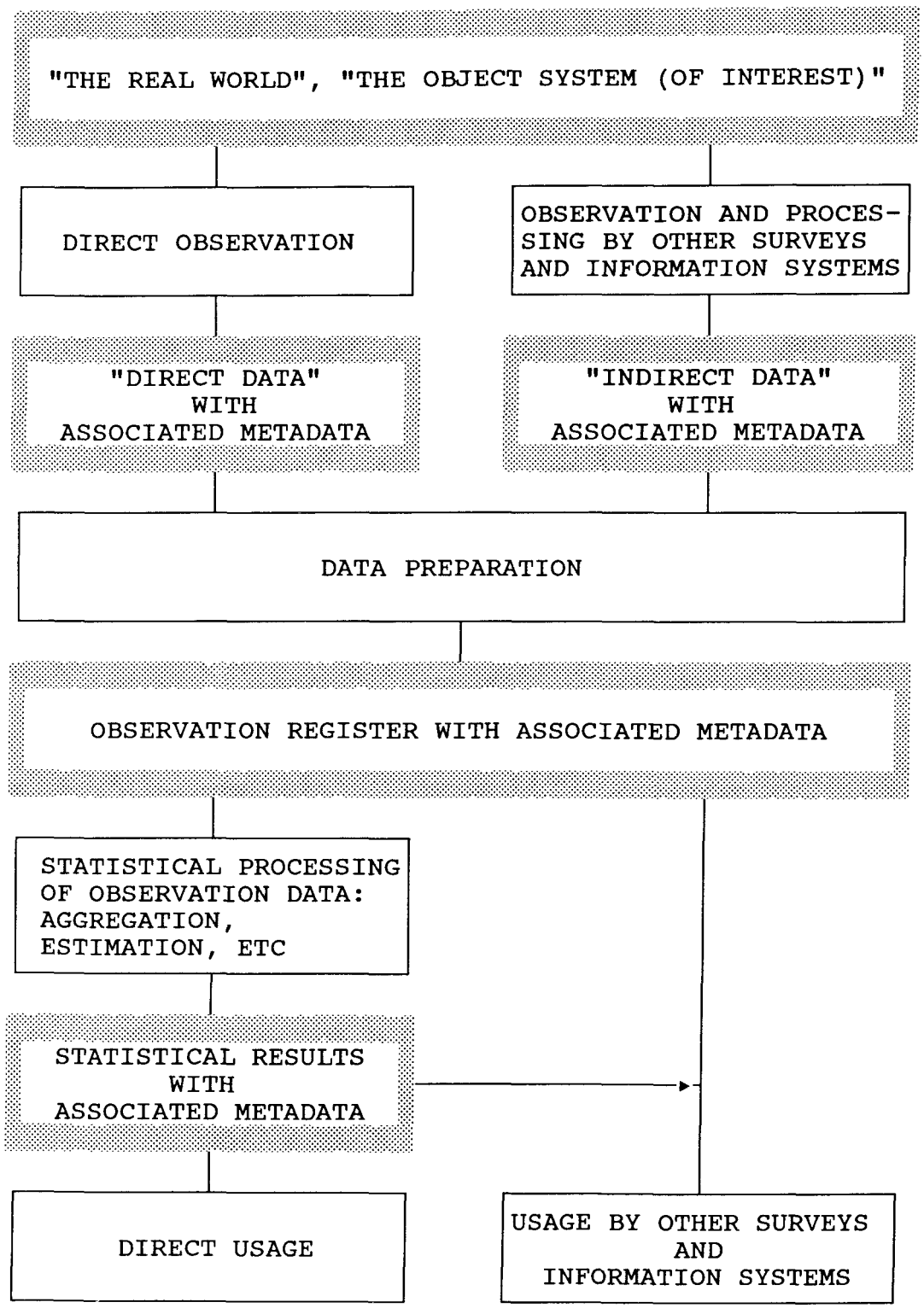


Figure 4.2b. *Survey production system: sequential architecture with explicit recognition of integrated data/metadata processing and links to other surveys.*

With the hardware/software technology available today, it has become natural to design a survey production system as a **database-oriented system**, that is, as a system whose different functions interact with one another and with a common database via a standardized **database interface** (for example SQL). The typical architecture of a database-oriented system for statistics production is visualized in figure 4.2c. In addition to the standardized database interface, it is nowadays common for such systems to be equipped with a **user interface** based upon standardized principles and software products (IBM CUA, Microsoft Windows, etc).

In figure 4.2c the database-orientation is limited to the production system of one, individual survey. Figure 4.2d indicates how the survey under consideration **shares** certain metadata (and object data) with other surveys.

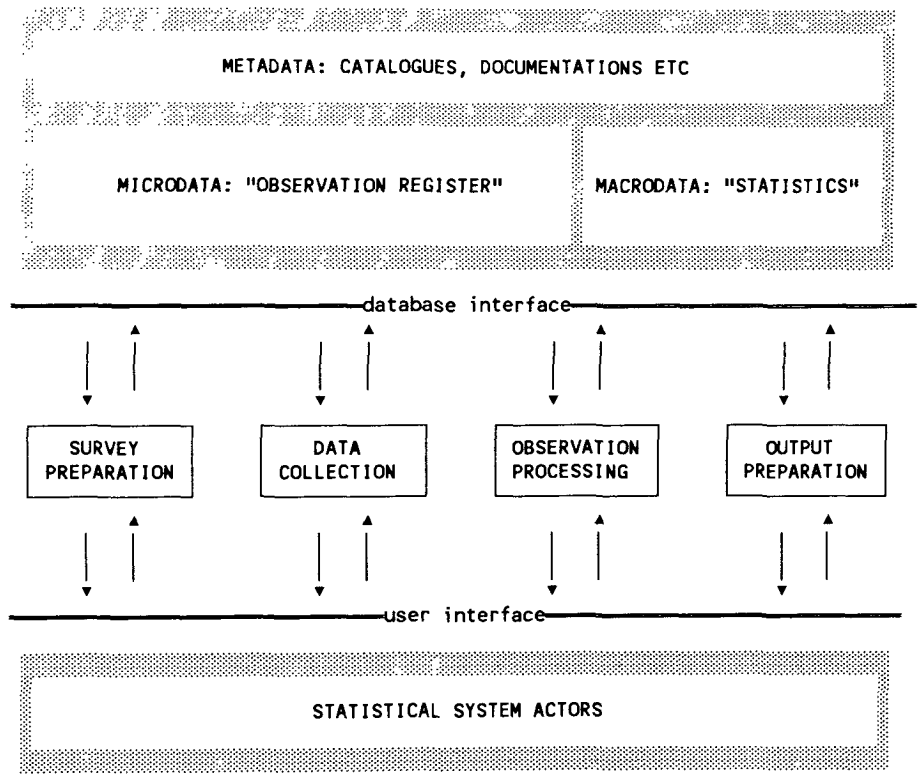


Figure 4.2c. *Survey production system: database-oriented architecture.*

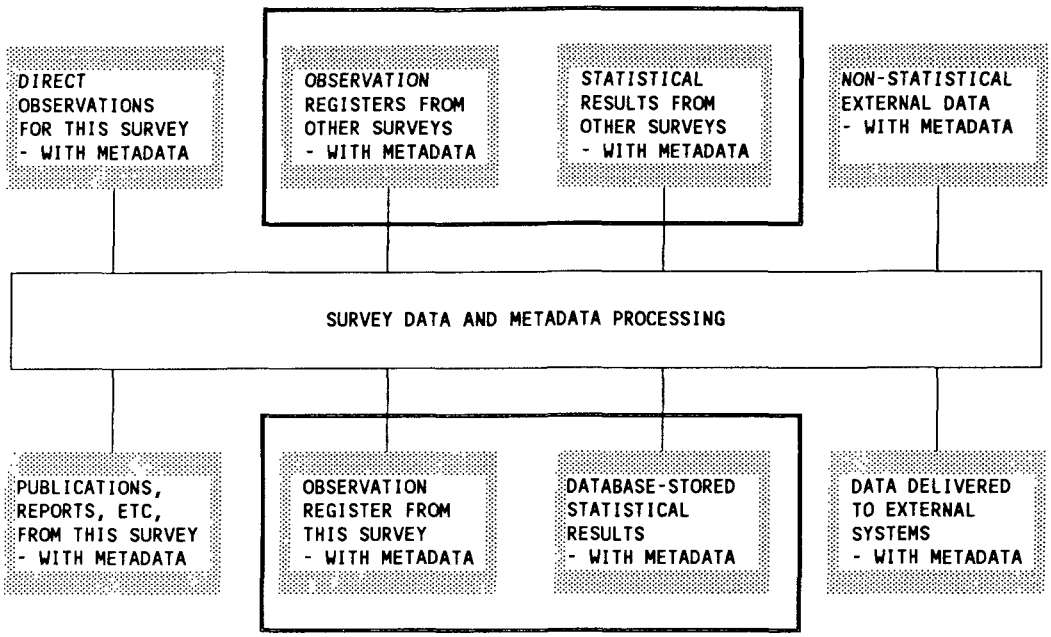


Figure 4.2d. *Survey production system: database-oriented architecture with emphasis on the sharing of data/metadata with other surveys.*

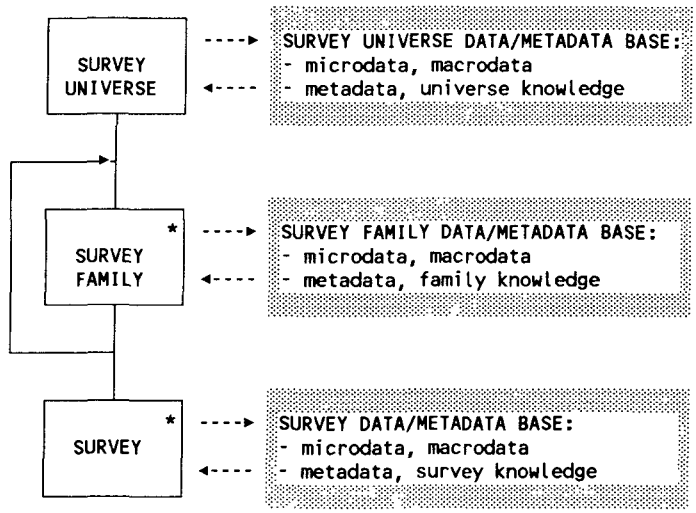


Figure 4.3. *Global, intermediate, and local level processes and databases of a universe of surveys.*

4.3 Data/metadata interactions between local, intermediate, and global levels

Figure 4.3, which is a revised version of a figure in [3], further illustrates the database orientation, and at the same time emphasizes the distribution of data/metadata and data/metadata handling processes over (at least) three different levels in the gross architecture of the data/metadata management infrastructure of a statistical office:

- the global, survey universe level;
- the intermediate, survey family levels;
- the local, survey occurrence level.

A well-functioning data/metadata infrastructure consequently requires well-designed communication paths between three or more levels of data/metadata storage and processing.

4.4 Data/metadata exchange between input-oriented and output-oriented systems

Most statistical offices are **input-oriented** in the sense that the statistical surveys, through which the statistical office collects its data, are also the natural "building-blocks" in its organisation.

A circumstance, which further strengthens the input-orientation of some statistical offices, is that there are relatively few "natural" and "hard" links between different survey production systems. The surveys are carried out largely independently of each other, and the data from the surveys are being compared and integrated only on the macro level, when the underlying surveys and survey production systems have already completed their tasks.

The relations between different surveys are somewhat different in a country like Sweden, where a relatively large number of data assets, notably those emanating from administrative sources, are being jointly used by many surveys. (By volume, administrative sources account for more than 95% of the input data to statistical surveys carried out by Statistics Sweden.) In such an environment one gets many "natural", "hard" links between survey production systems.

Other features that may stimulate logical and physical integration of survey production systems are the use of common sampling frames and positive (or even negative) coordination of the samples for different surveys.

Many statistical offices have recently committed themselves to become more **output-oriented** in the future. The planning and development of output-oriented databases (and metadata systems), serving the special needs and requirements of relatively well-defined external user groups, are steps in this direction.

Output-oriented database systems need to relate data from different surveys, and they often need to be equipped with special software and metadata tools, based on special methodology, for reconciling data from different sources, and for helping the users to interpret and analyze the data in adequate ways.

Databases containing time series of economical statistics as well as databases

with regionally structured data are common examples of early initiatives in the field of output-oriented systems. Since a relatively wide range of similar (and partly overlapping) systems can be foreseen to be proposed in the future, it is strategically important for statistical offices to consider how such initiatives, which should certainly be encouraged, could be best supported and coordinated by an appropriate data/metadata management strategy. We shall now briefly discuss this issue.

An **output-oriented database system** consists of two major components:

- the **database** component, containing data and metadata; and
- the **user interface** component, containing a user interface, supported by software and (additional) metadata.

The **database component** of an output-oriented database system could be physically located in at least three different places:

(a) **"output-locally"**, in connection with the user interface component;

(b) **"input-locally"**, in connection with each one of the surveys that have to provide the output-oriented database with data (and metadata);

(c) **"centrally"**, in a common database, **"the corporate database"** of the statistical office, which provides all output-oriented systems with the data (and metadata) that they need, and which in turn is more or less continuously updated with data from the underlying surveys.

With a growing number of output-oriented database systems, alternative (a) will lead to a very complex systems architecture, with a lot of complex communication (see figure 4.4ab) and physical duplication of data.

In addition, alternative (a) will lead to costly repetition of (more or less) the same systems development and maintenance activities, and (probably) a lack of uniformity in data and metadata management, which will in turn inevitably lead to a lack of uniformity in the interfaces between the statistical office and its information users.

Alternative (b) seems attractive from a theoretical point of view, but it requires a very sophisticated system for distributed database management. There are software systems, which claim to have such capabilities, but I would be surprised if there is (yet) a software product that would really satisfy the requirements of a statistical office, even if kept on a relatively modest level. The technical and conceptual problems involved are overwhelming.

Like alternative (a), alternative (b) implies a complex communication pattern between the output-oriented database systems and and the surveys providing the data and metadata; once again figure 4.4ab is applicable.

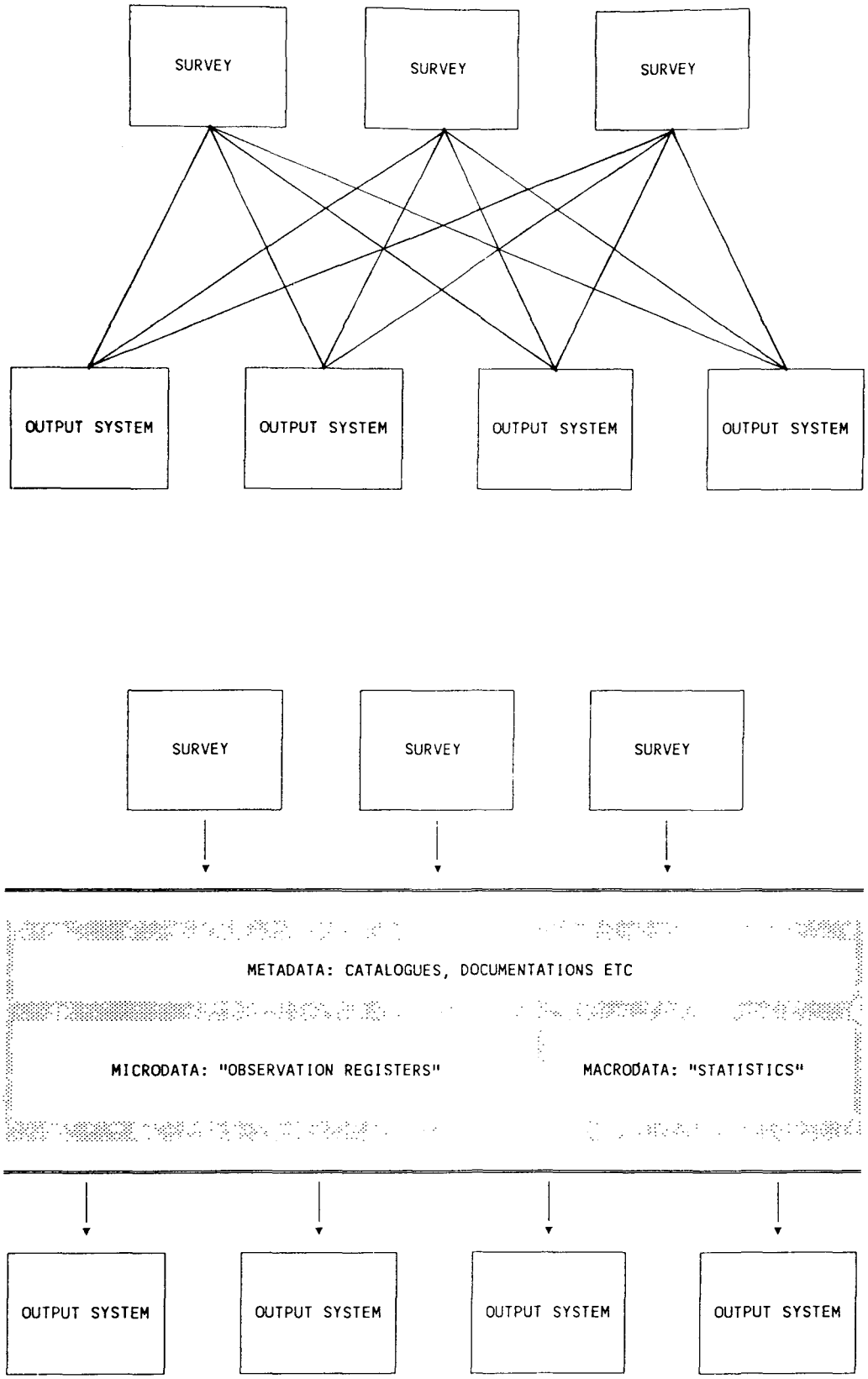


Figure 4.4. *Two alternative architectures for data and metadata management in a statistical office. The architecture in the lower part of the diagram (figure 4.4c) corresponds to "alternative (c)" in the text. The upper part of the diagram (figure 4.4ab) covers both "alternative (a)" and "alternative (b)".*

Alternative (c), with a central database component, minimizes the volume and complexity of the necessary communication between the surveys and the output-oriented systems; cf figure 4.4c. If m is the number of surveys, and n is the number of output-oriented system, the maximum number of communication interfaces will be (m + n) with this architecture, instead of (m * n) as with the other alternatives.

At least for the time being, it seems advisable that statistical offices adopt alternative (c) as the basis for its data/metadata management strategy, and that it should initiate and systematically carry out a number of well synchronized activities, in order to implement this data management strategy.

A data/metadata system based upon on architecture (c), as specified above, will be **"conceptually central"** in the sense that it will function as a "switch" or a "clearing-house" between the input-oriented survey systems for data collection, etc, and the output-oriented systems for retrieval, analysis, presentation, and distribution of statistical information, emanating from (often) several different survey sources.

Will a data/metadata system based upon architecture (c) also be **"physically central"**? At least in a long term perspective, this will not necessarily be the case. Theoretically, "the corporate database" could very well be "physically distributed", as long as all data and metadata can be conceptually interfaced in accordance with certain well-defined standards. However, such an architecture would have to be very sophisticated from a technical point of view, and robust, efficient solutions to these problems will hardly exist during the next few years. Moreover, a physically distributed architecture would require a discipline within the statistical office, for which most organisations of this kind are not yet ready.

Thus, for a foreseeable future, I think that the corporate databases of statistical offices will be physically central; at least this is a stage, which has to be reached first, before distribution and decentralization can be considered. However, there are several technical platforms available for a physically central corporate database, including mainframe-based and minicomputer-based solutions.

However, even a physically central corporate database of a statistical office should be able to interact, via data, metadata, and control flows, with more local systems, both on the input and on the output side.

4.5 An emerging target architecture with combined features

In sections 4.1 - 4.4 we have discussed a number of desirable features of the architecture of the data/metadata infrastructure of a statistical office:

- it should integrate data and metadata management;
- it should be database-oriented and support easy sharing of data and metadata within and between statistical systems;
- it should support data/metadata management on local, intermediate, and global levels, and facilitate communication between the levels;
- it should be able to handle relatively complex interactions between input-oriented and output-oriented statistical systems.

Figures 4.5 and 4.6 try to visualize an architecture, which satisfies this combination of desirable features.

Figure 4.5 basically looks upon the interaction between input-oriented and output-oriented statistical systems from the point of view of an individual, input-oriented survey. During the planning, operation (and primary use), and evaluation phases of such a survey, there are primarily interactions between the survey activities and the survey-related, local-level data/metadata base. For certain types of activities there will also be needs to make references to intermediate-level and global-level data/metadata bases, as was examplified in the scenarios in chapter 3 of this report. For example, it may be necessary to refer to classifications and variable definitions that are stored and maintained in metadata bases that are shared by a group of surveys, or by the organization as a whole.

In the lower part of figure 4.5 (below the dotted line) an alternative, output-oriented perspective is indicated in the form of three major activities of the "use" phase of an output-oriented retrieval, presentation, and analysis system. It should be noted that these output-oriented activities are running under quite different coordination and synchronization principles than the input-oriented survey activities. However, the two flows of activities are related in the sense that the output-oriented system now and then makes (secondary) use of data and metadata originating from the particular (input-oriented) survey under consideration. They do so by making references to some global-level, intermediate-level, or local-level data/metadata base, to which the survey has, at some stage, contributed.

Figure 4.6, if studied from the bottom and upwards, looks upon the interaction between statistical systems from the point of view of an output-oriented retrieval and analysis system, that is, from an end-user relevant perspective. Such a system primarily interacts with a global data/metadata base, or possibly with a subset of such a data/metadata base, tailored to the needs of a particular user category, to which the particular end-user belongs. Occasionally the system will have to make references back to some intermediate-level data/metadata base, or even to some local-level data/metadata holding.

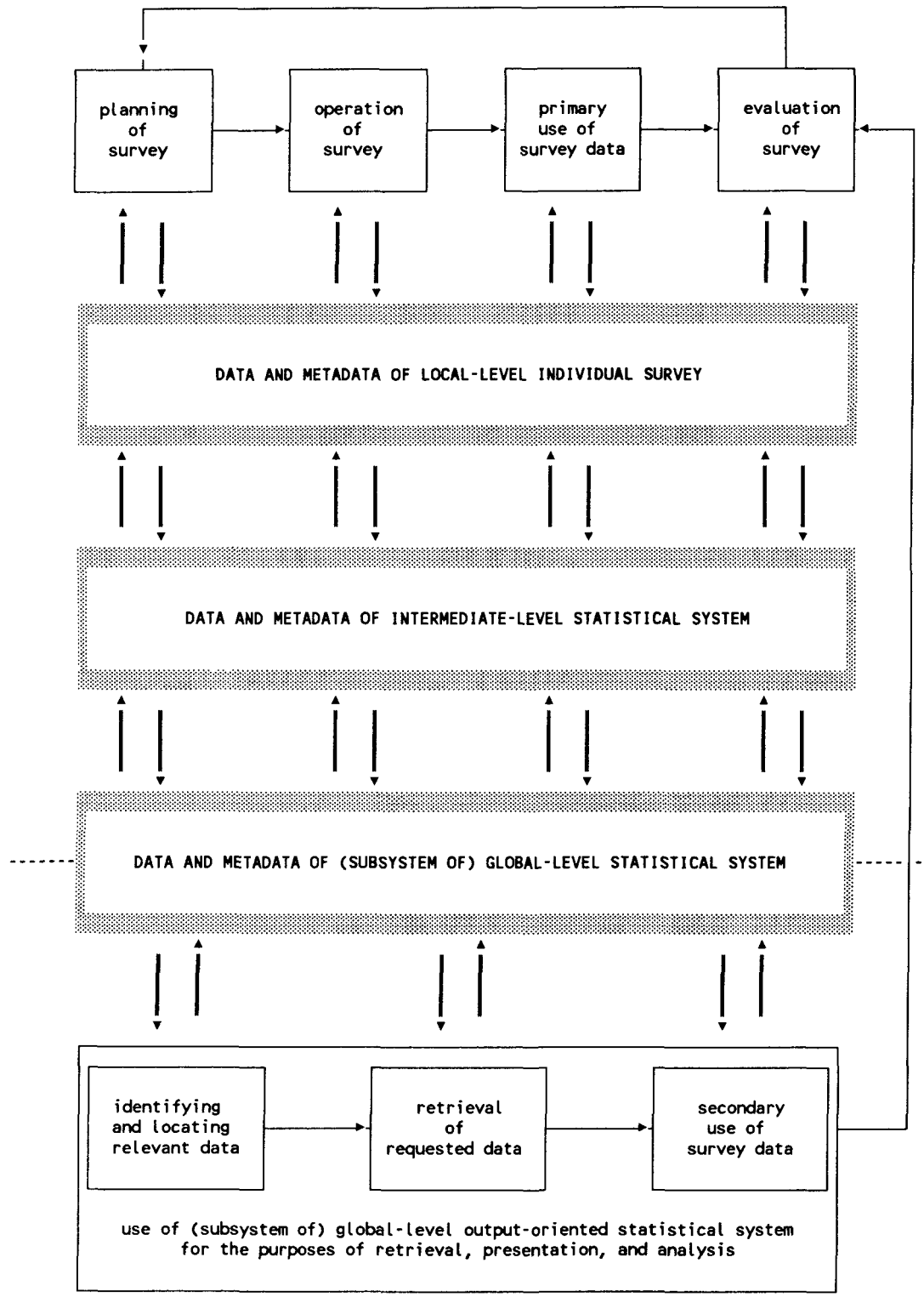


Figure 4.5. *The interaction between input-oriented and output-oriented statistical systems, seen from an individual (input-oriented) survey's point of view. The fat arrow lines indicate data/metadata flows, the thin ones control flows.*

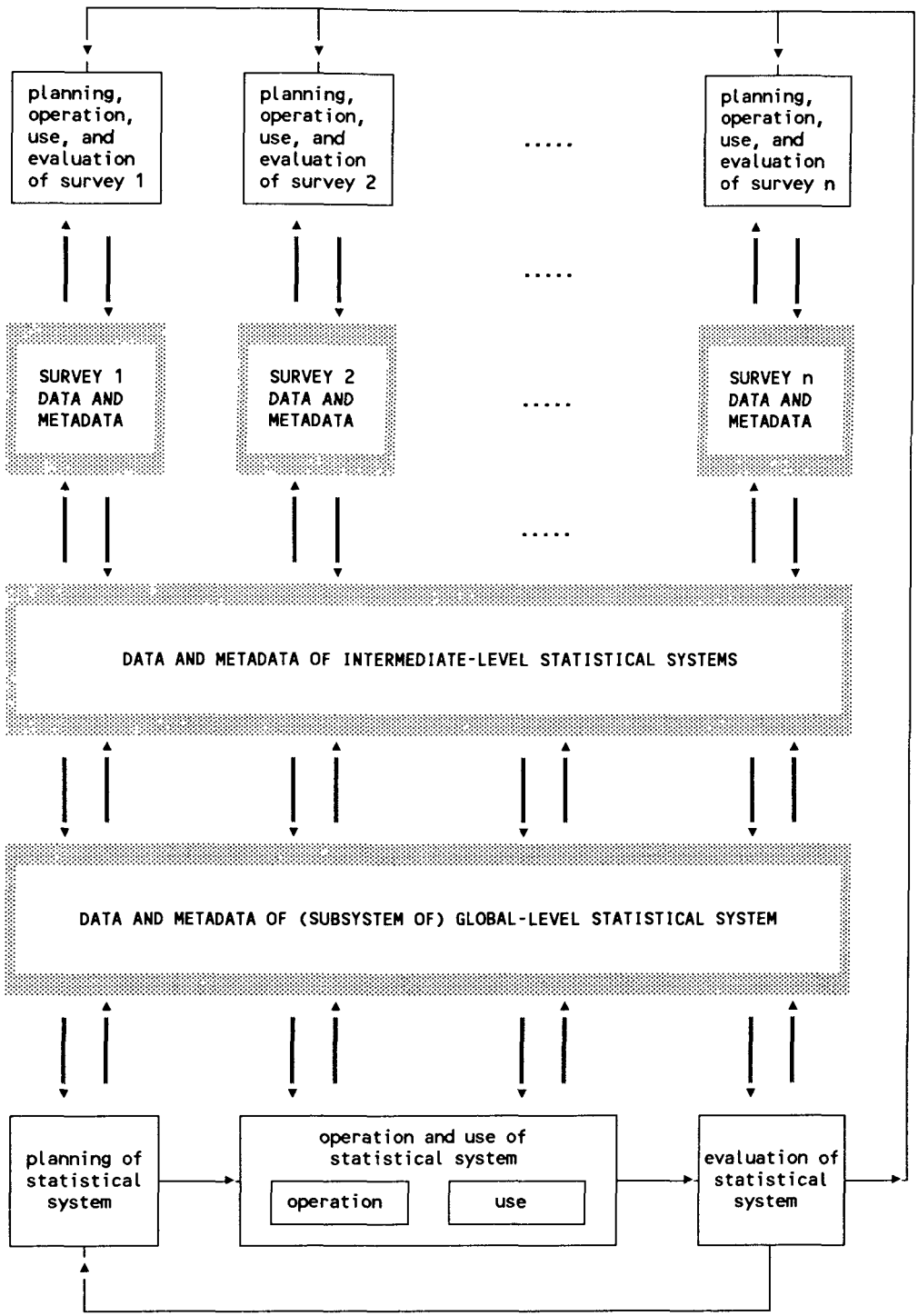


Figure 4.6. *The interaction between input-oriented and output-oriented statistical systems, seen from an output-oriented retrieval and analysis system's point of view.*

4.6 Harmonizing the metainformation infrastructure with other infrastructures of a statistical office

The metainformation infrastructure is not the only important infrastructure of a statistical office. It is not even the only infrastructure in the field of information systems. We have already in this chapter emphasized the necessity to integrate the infrastructures of metadata and (object) data management. Another information systems infrastructure, which rapidly gains importance in many statistical offices, is the **Office Information System (OIS)** of the statistical office, typically containing components for functions like word processing, electronic mail, document storage and retrieval, etc.

The metainformation infrastructure can certainly benefit in several ways from being harmonized with the OIS. One rather trivial, and yet important reason is that metadata management itself requires a lot of word processing and document handling. Staff members of statistical offices will appreciate - or rather expect and require - that functions which are common for OIS and metadata management can be done in the same way, and with the same tools, within both infrastructures.

5 References

[1] Bengt Rosén & Bo Sundgren: *"Documentation for reuse of microdata from the surveys carried out by Statistics Sweden."* Statistics Sweden 1991. Original report in Swedish. English translation available.

[2] Bo Sundgren: *"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*. The report is also available from Statistics Sweden as R&D Report 1991:9.

[3] Bo Sundgren: *"Statistical metainformation and information systems."* Report for the October 1991 Meeting of the UN/ECE METIS Group, established within the programme of work of the Conference of European Statisticians. The report is also available from Statistics Sweden as R&D Report 1991:11.

[4] Bo Sundgren: *"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.

[5] Bo Sundgren: "Conceptual modelling as an instrument for formal specification of statistical information systems." ISI 47th Session, Paris 1989. The paper is also available from Statistics Sweden as R&D Report 1989:18.

[6] Bo Sundgren: *"Some properties of statistical information: pragmatics, semantics, and syntactics."* Statistics Sweden 1991.

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

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.