Statistiska centralbyrån Statistics Sweden

Statistical Metainformation and Metainformation Systems

Bo Sundgren

R&D Report
Statistics Sweden
Research - Methods - Development
1991:11

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 1991:11. Statistical metainformation and metainformation systems / Bo Sundgren. Digitaliserad av Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-1991-X101OP9111

Statistiska centralbyrån Statistics Sweden

Statistical Metainformation and Metainformation Systems

Bo Sundgren

R&D Report
Statistics Sweden
Research - Methods - Development
1991:11

Från trycket | Augusti 1991
Producent | Statistiska centralbyrån, utvecklingsavdelningen
Ansvarig utgivare | Åke Lönnqvist
Förfrågningar | Bo Sundgren, 08-783 41 48

ISSN 0283-8680
Printed in Sweden

STATISTICAL METAINFORMATION AND METAINFORMATION SYSTEMS

0 Introduction and summary

"Statistical metadata - METIS" is a newly established item (12.2.2:iii) in the programme of work of the Conference of European Statisticians (CES) within the United Nations Economic Commission for Europe (UN/ECE). A Joint Group - here called the *UN/ECE METIS Group* - will be formed with members from statistical offices in the ECE region. The present report has been written on request by the ECE Secretariat as a basis for discussion at the first meeting on METIS, scheduled to take place in the beginning of October 1991. According to the ECE request, the report should be a preparation of a first design of a pilot meta-information system, comprising

(1) analysis and evaluation of metadata in statistical information systems;

(2) design of the methodological framwork for a pilot metainformation system;

(3) elaboration of a pilot metainformation system;

(4) future work to be done.

The report consists of five main chapters. In **chapter 1** a number of concepts from information systems theory are defined and related to the issues of metainformation and metainformation systems.

Chapter 2 analyzes the purposes of metainformation and metainformation systems. Three main categories of users and usages are identified: clients (or end-users) of information systems, information system administrators (including development, maintenance, and operating staff), and software artifacts.

In **chapter 3** the concept of a pilot metainformation system for statistical environments is discussed, and a tentative definition is given for the purposes of the ECE/METIS Group. Some lessons and experiences from similar projects in the past are also mentioned.

Chapter 4 is the major part of the report. It contains an outline of an architecture for a statistical metainformation system. The architecture is based on the principle

that the structure of an information system should reflect the structure of the object system that the information system processes information about. Thus a statistical metainformation system should have a structure reflecting the structure of a statistical survey. Typical structures of statistical surveys and of systems of statistical surveys, so-called statistical systems, are described verbally, and with diagrams.

Two main sources of survey metadata are identified: survey design decisions and the survey process itself. Survey metadata (like other data) should be captured as close to the sources and as automatically as possible. In practice this means that survey metadata should be captured and organized as an automatical side-effect of other tasks during the planning, operation, and evaluation of the survey.

An input/output/process scheme of analysis is used for finding out more details about the metadata (and object data) processing that takes place during each one of the three major phases of a survey life cycle.

For statistical systems (survey universes and survey families) a decentralized approach with adequate coordination on each level is foreseen. Thus microdata, macrodata, and metadata have to be distributed between the levels of survey universe, survey family, and individual survey in such a way as to minimize redundancy and maximize autonomy on each level.

Chapter 5 proposes seven major areas for future work within the UN/ECE METIS Group:

- analyze metainformation needs for a statistical system, given requirements of - among other things - infological and procedural completeness, and including factual as well as rule-based metainformation;
- find "natural" metadata sources and "convenient" metadata collection procedures, taking into account the idea of generating as much of the metadata as possible by side-effects from other processes;
- establish a reference conceptual model for statistical metainformation;
- establish a reference flow model for statistical metainformation;
- establish some important metadata interfaces for statistical systems, taking into account other ongoing international standardization work, notably the UN/EDIFACT standardization efforts;
- make systematical studies of ongoing statistical metainformation projects in statistical offices and statistical organizations;
- experimental design and implementation of software and other tools for (components of) statistical metainformation systems.

1 Basic concepts

Note. The whole or parts of this chapter may be skipped by readers, who are already familiar with information systems theory. Even other readers may skip it at first reading, and return to it as needs arise during reading the rest of the report.

1.1 Information and data

Information and data are two sides of the same coin. **Information** is the **knowledge contents** of data, as interpreted by a human being. **Data** are **representations** of information on (in) some kind of physical medium. Computers can store and process information only in terms of data, and human beings can (probably) communicate information to each other only in terms of data.

1.2 The object of information and data

Information is always information *about something*, the **object of information**. A piece of information informs about a **piece of reality** of some kind; the piece (or aspect) of reality is then the object of the piece of information.

Since data represent information about something, data can also be said to be data about something, the **object of data**.

1.3 Metainformation and metadata

Metainformation is **information about information** (including information about data representations of information). Thus the object of metainformation is information.

Metadata is data **representing metainformation.**

In discussions concerning metainformation and metadata it may sometimes be difficult to maintain a strict distinction between the two concepts of "meta-information" and "metadata". In such situations it may be better to choose one of them (say: metadata) as the main concept (and term) of the discussion. If we choose "metadata" as the concept in focus, contents-oriented parts of the discussion may be carried out in terms of (contents-oriented) *aspects* of the metadata concept. This convention will be applied in later parts of this report (see chapter 4).

1.4 Information systems

1.4.1 Purpose

An **information system** is a system, which **supports decision-making** concerning some piece of reality, the **object system**, by giving the decision-makers access to information concerning relevant aspects of the object system and its environment.

The term "object system" turns out to have several connotations, if one subjects it to a sharper analysis. Here are some of them:

- The **object system of control**, or **domain of control**, is the piece of reality upon which the decision-makers act (directly).

- The **object system of interest**, or **domain of interest**, is the piece of reality about which the information system provides information, in order to facilitate decisions and actions visavi the object system of control.

- The **object system of action**, or the **subject system**, is the system of decision-makers and other actors in activities being served by the information system.

In practice, these and other interpretations of the object system concept are to a great extent overlapping. However, not least in statistical applications, the distinctions may sometimes be important. Consider, for example, the System of National Accounts (SNA) of a country. This information system supports high-level political decisions concerning the economical politics of a government. At least according to some economical theories, these decisions should primarily concern certain strategical, macrolevel parameters, such as money supply, interest rates, taxes, etc. But in order to arrive at these decisions, the decision-makers will need information based on numerous microlevel observations concerning individual companies and citizens of the country. Thus, although the object system of control in this case is essentially a macrolevel system, with a relatively small number of macrolevel objects and variables, the object system of interest contains a lot of microlevel objects and variables. The subject system would in this case include politicians (from the government and the opposition) and other actors (for example interest organizations), who participate in the formation of an economical policy of a country.

1.4.2 Functions

An **information system** typically contains such **functions** as

- **collection** of information;
- **storage** of information;
- **transformation** of information;
- **communication**, **retrieval**, and **distribution** of information;
- **user interaction** with information.

The **architecture of an information system** could be such that it contains a subsystem for every major function, for example,

- an **input-oriented subsystem** for the information collection function;
- an **information base**, or **knowledge base**, for the information storage function;
- a **knowledge transformation** subsystem, or **inference engine**, for the information transformation function;
- an **output-oriented subsystem** for communication and distribution of information;
- a **user interface** for interactions between different categories of users and the information system.

An information system works in terms of **information processes** that process (input) information and produce (output) information. Thus an information system (as well as an information system function/subsystem) consists of a system of information processes, using and producing information.

1.4.3 Implementation aspects: data (processing) systems

Like information is physically represented by data, an information system is physically represented, or implemented, by a **data (processing) system**. Each function/subsystem/component of an information system has a data (processing) system counterpart; in particular information processes are physically represented by data processes, using and producing data.

1.5 Metainformation systems and metadata systems

A **metainformation system** is an information system that uses, stores, and produces metainformation for the purpose of supporting decision-making concerning an information system. Thus the object system of a metainformation system is an information system.

A metadata system is the data (processing) system counterpart of a metainformation system. Thus a metadata system is a data (processing) system that uses, stores, and produces metadata.

A metainformation system can be more or less (logically and physically) integrated with the information system that it uses, stores, and produces (meta)information about. In one extreme, a metainformation system can be completely separate from its information system counterpart(s); all linking of metainformation with object information then has to be done by human intervention. In the other extreme the linking is done automatically as far as is logically possible.

1.6 Facts and rules

There are several different semantical forms of information. Two common forms of information in information systems are facts and rules. Factual information can be formalized in terms of messages (see 1.7 below). Rule-based information can have the form of, for example, a definition, a law, an algorithm, or a description of (typical) behaviour.

Most information processes in a (statistical) information system use and produce factual information; however, the processing as such is typically controlled by rule-based information. In advanced information systems some processes may use and produce rule-based information, in addition to being controlled by such information.

Traditionally, the subdiscipline within information systems theory that is known as **artificial intelligence** has focused its interest on rule-based information, whereas other subdisciplines, such as **administrative data processing** have focused on factual information. Today it seems most appropriate to recognize both types of information within one and the same theoretical framework, and that is the approach that we shall try to apply in this study of metainformation and meta-information systems.

1.7 Formalized modelling of information

1.7.1 Factual information: messages and message types

Factual information can be formalized in terms of messages. A **message** is built up

from **references** to entities in a piece of reality that is of interest, a so-called object system (cf section 1.4.1 above). Complex messages can be broken down into simpler messages, until the level of **elementary messages**, or **e-messages**, is reached. E-messages are built up from references to four kinds of object system entities:

- **objects**;
- **properties**;
- **relations** (between objects);
- **times** (points of time, as well as time intervals).

There are two typical e-message structures: property type e-messages, and relational e-messages. An **e-message of property type** informs about the fact that a certain object has a certain property at a certain time:

(1.1) $<\rho(o), \rho(p), \rho(t)>$;

where

- $\rho(o)$ is a reference to an object in the object system;

- $\rho(p)$ is a reference to a property in the object system; many properties are referred to in terms of a variable (sometimes called an attribute) and a value according to the formula: $<\rho(V) = \rho(a)>$;

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

An **e-message of relational type** informs about the fact that n objects are related to each other in a certain way at a certain time:

(1.2) $<<\rho(o_1), ..., \rho(o_n)>, \rho(R), \rho(t)>$;

where

- $\rho(o_1), ..., \rho(o_n)$ are references to objects in the object system;

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

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

The object system structural counterparts to information system e-messages are called **e-constellations**. In analogy with property type e-messages and relational type e-messages, there are **property type e-constellations** and **relational type e-constellations**:

(1.1') $<o, p, t>$;

(1.2') $<<o_1, ..., o_n>, R, t>$;

When modelling object systems and information systems it would not be practical work in terms of *individual* e-constellations and e-messages. Instead one defines

such models in terms of e-constellation *types* and e-message *types*.

The structure of a **attributive e-message type** is:

(1.3) $<\rho(O), \rho(V)>$;

where $\rho(O)$ is a reference to an object type, and $\rho(V)$ is a reference to a variable.

The structure of a **relational e-message type** is:

(1.4) $<<\rho(O_1), ..., \rho(O_n)>, \rho(R)>$;

where $\rho(O_1), ..., \rho(O_n)$ are references to object types, and $\rho(R)$ is a reference to an n-ary relation.

1.7.2 Rule-based information: logic and programming languages

For rule-based information there are formalization tools readily available in the disciplines of logic and of programming languages.

1.8 Formalized modelling of metainformation

1.8.1 Metainformation e-messages

Since the object system of a metainformation system is an information system, metainformation e-messages inform about objects, which are parts of an information system. Maybe the most obvious metainformation objects are e-messages and their components, but, in principle, all entities, about which we need information, in order to be able to interpret and process information in an information system, are to be regarded as potential (meta)objects in the metainformation system concerning the information system.

When we have identified the (meta)objects of a metainformation system, we must find (meta)relations and (meta)variables in much the same way as when we develop a conceptual model for an "ordinary" information system. The conceptual model of a metainformation system can be illustrated by an object graph, in the same way as we can visualize the conceptual model of an "ordinary" information system.

The exact contents of the conceptual model of a metainformation system will be determined by the purposes of the metainformation system, as we shall discuss in the next chapter of this paper.

1.8.2 Rule-based metainformation

Rule-based metainformation can take such shapes as

- definitions of concepts in the object system, expressed as first-order predicate logic formulae;

- logical formulae expressing relations between different concepts or "laws of conduct" for object system entities;

- application programs describing object system processes;

- editing and coding rules.

1.9 Statistical information and statistical information systems

The most typical feature of a statistical information system is that it contains **aggregation processes**, that is, processes that transform sets of so-called microlevel e-messages into so-called macrolevel e-messages. A **microlevel e-message** is an e-message concerning an individual object in a collective of objects. A **macrolevel e-message** is an e-message about the collective as a whole, or about a subcollective of this collective. Macrolevel e-messages are derivable from microlevel e-messages about the individual objects in the collective. and about subcollectives of the collective:

(1.5) $F(\{\langle \rho(o_i), \rho(V)=\rho(v_i), \rho(t_i)\rangle \mid o_i \in O\}) = \{\langle \rho(O_j), \rho(W)=\rho(w_j), \rho(t_j)\rangle \mid O_j \subset O\}$

F is the **aggregation function**, O is a collective or **population (of interest)**, o_i (i = ...) are **individual objects** in the population, V is a **(microlevel) variable** that is relevant for the objects in the population, and W is a **characteristic** or **(macrolevel) variable** of the population O and some domains of interest O_j within the population.

Microlevel e-messages are also called **microinformation e-messages**. They are **input e-messages** to an aggregation process.

Macrolevel e-messages are also called **macroinformation e-messages**. They are **output e-messages** from an aggregation process.

According to the principle that *"one man's ceiling is another man's floor"*, macroinformation from one aggregation process may be regarded as microinformation by another aggregation process.

The term **"statistical information"** (and **"statistical data"**) is sometimes used for **macroinformation (macrodata)** *only*. However, it is at least equally common to include *also* **microinformation (microdata)** in the meaning of the term. The latter practice will be followed in this paper. Thus "statistical information" will denote both (statistical) microinformation and (statistical) macroinformation, and "statistical data" will denote both (statistical) microdata and (statistical) macrodata.

1.10 Statistical metainformation and statistical metainformation systems

Statistical metainformation is information about statistical information. Statistical metadata are data representations of statistical information.

A statistical metainformation system is an information system that processes, stores, and produces statistical metainformation.

Statistical information systems and statistical metainformation systems can be more or less closely linked to each other, that is, more or less integrated.

2 Purposes of metainformation and metainformation systems

2.1 Supporting decision-making concerning information systems

The purpose of an information system in general is to support some kind of decision-making visavi a piece of reality, the object system; (cf section 1.4.1 above). There may be a wide variety of decision types that are supported by an information system, ranging from routine decisions, taken more or less mechanically by humans and/or computers, to strategical decisions, taken under unique circumstances and after long reflection and discussion.

From this definition of the purpose of an information system in general, we may derive the purpose of a metainformation system. Since the object system of a metainformation system is itself an information system, the purpose of a meta-information system will be to support decision-making visavi an information system.

2.2 Types of decisions and decision-makers

Which kinds of decisions concerning information systems could be supported by a metainformation system, and who are the typical decision-makers?

We may distinguish between **three major categories** of decisions and decision-makers to be supported by a metainformation system: clients (end-users), administrators (including development and maintenance staff), and software artifacts.

2.2.1 Clients

Firstly, among the decision-makers using a metainformation system, we have the **clients** of the information system, which is the object system of the metainformation system. The clients are often called **end-users** or just **users**; however we shall avoid this term, since it may be ambiguous. The clients of an information system will need to know such things as

- Which information is available from the information system?
- Is this or that particular information type available?
- How is this or that concept defined?
- What is the quality of the information?
- Can I have the information processed or analyzed in a certain way?
- How do I put my request to the information system?
- How much would it cost to satisfy my request, and how long would it take?

Most of the (meta)information needs belonging to this category will be of an **infological** nature, that is, they will concern the meaning and contents of information and information system functions, rather than **datalogical** matters concerning data representations and technical solutions; however, as can be seen from at least one of the examples in the list above (How do I put my request...), even the clients may have some needs for datalogically oriented metainformation.

2.2.2 Administrators

Secondly, among the decision-makers using a metainformation system, there are

what we may call **administrators** of the information system, people for whom the information system in itself is a *task* rather than a *means to an end*. This category includes those who develop and maintain the information system and its different components, those who operate the information system and give assistance to the clients, etc.

To the extent that the administrators perform client functions, or assist clients to perform such functions, their (meta)information needs are, of course, similar to those of the clients. Tasks like information systems development and maintenance are associated with more unique (meta)information needs.

Administrators will need both infologically and datalogically oriented meta-information. A good information systems documentation will give many examples of metainformation needed by those who develop and maintain an information system: conceptual definitions, system/subsystem/component structures, system flows, process descriptions, data descriptions, record layouts, etc.

2.2.3 Software artifacts

Thirdly, there is a category of information system "decision-makers" and meta-information users (or rather metadata users) that consists of **software artifacts**, developed by human beings, but executed by computers. All software products use metadata. However, traditionally these metainformation needs have not been very explicitly recognized, and metadata and metadata (sub)systems have not been consciously designed and organized to the same extent as (object) data and (object) information (sub)systems. Neither has metadata handling been systematized and automated to the same extent as (object) data handling.

Traditionally the metadata used by software products were part of the programs themselves, or appeared as manually entered parameters and control statements. With the advent of database ideas, including concepts like data/program independence, the metadata aspects of software and information systems design became much more explicitly recognized, and metadata handling began to be systematized and automated. This development has continued over the years, and the CASE tools (CASE = Computer Assisted Systems /or Software/ Engineering) launched today can be seen as yet another step of progress.

Obviously, software artifacts in the first place use datalogically oriented metadata. However, there are also some examples of infologically oriented metadata for this category of metadata users: conceptual schemas used by database management systems, design aids for conceptual modelling in CASE tools, etc. In the future we are likely to be offered software products, where all client/system communication and most administrator/system communication is in terms of infologically oriented concepts. Naturally such products will have to have rather sofisticated metadata handling, including both infological and datalogical metadata - and complex transformations forwards and backwards between infological and datalogical metadata.

2.3 Purposes of statistical metainformation and metainformation systems

So far the discussion of the purposes of metainformation and metainformation systems has been quite general and not limited to statistical environments. If we

limit the analysis to *statistical* metainformation and *statistical* metainformation systems, we can, on the one hand, be more precise and concrete, and, on the other hand, we must then consider more deeply certain typical features of statistical information and statistical information systems.

The three major types of usages of metainformation, identified in section 2.2 above, will, of course, appear also in statistical environments. Thus clients, administrators, and software artifacts, as just defined, will appear as users of statistical metainformation and metainformation systems. What can we add to the general descriptions of these users, when we consider them in a statistical environment?

A **client** of a statistical (meta)information system will often start out from a rather vague notion of his or her information needs. This comes from the fact that the decisions supported by statistical information are typically of a **directive** (strategical) rather than of an **operative** nature. In an operative decision environment it is usually a "yes/no" decision to determine whether certain information is needed: either the information is need, or it is not needed.

In a directive decision environment some kind of **cost/benefit analysis** is required, at least in principle, in order to determine whether certain information is needed (or rather: would be worth its cost) or not. The guiding principle for the cost/benefit anlysis should be the following one:

> *Cost/benefit principle for directive information*
>
> Let ΔB be the marginal increase in the value of the (improved) quality of the decision, provided that a certain piece of information I is available to the decision-makers, and let ΔC be the extra costs for producing or retrieving I. Then I should be produced/retrieved if and only if ΔB exceeds ΔC. (It is worth noting that ΔC should include all kinds of costs, for example the costs in terms of lower decision quality, if the decision is delayed as a result of the production/-retrieval of I.)
>
> *End of cost/benefit principle for directive information*

In practice, a client of a statistical information system is hardly likely to make explicit cost/benefit analyses as formulated by the principle above. But implicitly the client is very likely to apply the principle when approaching the statistical information system, considering what information he or she will request. It is easy to imagine several metainformation features that could help the user to make information requests that are as rationally grounded as possible. In addition to (meta)information concerning database contents and retrieval and processing possibilities and costs, one could mention more spectacular metainformation features like simulation systems for trying out possible information requests and analyses on the basis of some small-scale version of the complete information system.

The typical **administrator** of a statistical (meta)information system is an employee of a statistical organization like a national statistical office or (a part of) an international organzation. Administrators are typically statisticians, subject-matter specialists, EDP staff, or some combination of these categories. They are typically

responsible for designing, developing, operating, and maintaining surveys within a certain area of (subject-matter) competence.

Survey knowledge bases and expert systems

Many statistical surveys are repeated at regular intervals, for example monthly, quarterly, or yearly. For such **survey series** it seems appropriate for the administrators to maintain a metainformation system that will have the function of an **expert system** for the administrators themselves. Such a metainformation system could be called a **survey knowledge base**, and it would be instrumental in codifying existing practices and training new staff members. The survey knowledge base would also be a useful basis for producing the metainformation that needs to accompany the (macrodata) statistics produced by the survey as well as the survey (micro)data to be stored for future (re)usage.

To the extent that survey knowledge bases could be read and "understood" by computers they could also be used for feeding the third category of statistical metainformation users, the **software artifacts**, with the metadata that they need. The software artifacts could be either application software, like statistical analysis packages, or (CASE) tools for producing applications from high-level specifications.

2.4 Infologically and procedurally complete information systems

One major purpose of the metainformation in an information system is to support the client users so that they will be able to interpret all output data from the system in a "reasonably" correct way. An information system, which contains all the metainformation needed for this purpose as an integrated part of the information system itself, is in a certain sense **self-contained**; we shall also say that it is **infologically complete** with respect to the information that it contains (or can produce), and with respect to the client users that it supports.

Another major purpose of the metainformation in an information system is, as we have discussed in section 2.2 above, to support software artifacts and information system administrators with the information that they need in order to operate and maintain the procedures of the information system. An information system, which contains all the metainformation needed for this purpose as an integrated part of the information system itself will be called **procedurally complete** with respect to the functions of the information system.

3 A pilot metainformation system for statistical environments

3.1 The concept of a pilot metainformation system

One of the tasks of the ECE/METIS group is to develop a so-called **pilot metainformation system.** The meaning of the term is not very clear. To begin with, it raises the question

(3.1) What is a pilot system?

According to a dictionary "pilot" as an adjective could mean

(3.1a) "serving as a guiding or tracing device"; or

(3.1b) "an activating or auxiliary unit"; or

(3.1c) "a trial apparatus or operation".

The interpretations (3.1a) and (3.1c) seem to be most relevant in our present ECE/METIS context. However, it remains to decide what nature the pilot system should have. Should it be an implemented, computerized system, let be on an experimental basis, or should it be "only" a well documented design of such a system? Under all circumstances a good design is necessary, so it seems wize to start with that task. Then it could be discussed which systems, subsystems, and/or components that may be worthwhile to develop, and how the development should be organized and technically solved.

Then we come to the question

(3.2) What kind of system is it that the pilot system should be a pilot system for?

With the background lined out so far in this paper, and with the assumption that it is a pilot *statistical* metainformation system that we are talking about, the following interpretations seem possible:

(3.2a) an integrated metainformation subsystem of the information system corresponding to a (typical) statistical survey;

(3.2b) an autonomous metainformation system for a (typical) statistical survey (series), containing components like an expert system and a knowledge base for the administrators and clients of the survey;

(3.2c) an autonomous (or partially integrated) metainformation system for a collection of related surveys, or a database emanating from such a collection of related surveys; in the extreme case "the collection" could cover all surveys conducted by the statistical organization under consideration, maybe even augmented by some important, related surveys conducted by other organizations;

It seems clear that statistical organizations in the future will need to have well designed metainformation systems of all three types mentioned above. As a first step one should therefore maybe think of the pilot metainformation system as a

(pilot) system of (pilot) metainformation systems.

3.2 Tentative definition of the ECE/METIS pilot metainformation system

At this stage it seems appropriate to make a first attempt to formulate the task(s) for the envisaged ECE/METIS pilot metainformation system:

3.2.1 Gross architecture of a system of metainformation systems

The gross architecture should primarily be formulated on a conceptual level, but technological problems and possibilities of strategical importance should also be addressed, for example, which types of software, data organizations, and hardware (processors, storage media etc) should be explored for major functions in the pilot metainformation system.

On a high level, the gross architecture will consist of subsystems of the three types indicated in section 3.1, that is, metainformation (sub)systems of types (3.2a), (3.2b), and (3.2c), possibly together with metainformation (sub)systems of some other types.

3.2.2 Identification of subsystems and components of particular interest

Metainformation subsystems and components (within the gross architecture indicated in 3.2.1) of particular interest for the statistical offices and organization that are cooperating within the ECE/METIS project, should be identified by the ECE/METIS group.

3.2.3 Detailed conceptual design of selected subsystems and components

For the selected subsystems and components identified in 3.2.2 a more detailed conceptual design should be worked out. To a certain extent such designs are available from the previous SCP/METIS project, which in such cases could be used as a basis for further design work.

A reasonable ambition level for the work to be done under this item could be

- to develop standardized conceptual designs for certain metainformation subsystems/functions/components, which could be used as a starting point for those statistical offices and organizations who themselves want to design an implement metainformation (sub)systems;

- to develop a conceptual framework of reference for certain metainformation subsystems/functions/components, which could be used for description, evaluation, and comparison of such subsystems etc, which have actually been developed by statistical offices or elsewhere.

3.2.4 Experimental implementation of selected subsystems and components

As always, it could be debated whether it is at all meaningful and rational for international groups (like ECE/METIS) to actually develop and implement software. *If* a decision is taken by the ECE/METIS group to actually start such development, the following conditions should be ensured:

- the development should be based on a detailed and very well documented conceptual design;

- the selected subsystem/component should not be too big and/or complex, and it should be relatively independent of other subsystems/components, or at least have very well defined and documented interfaces to all related subsystems/components;

- one statistical office should have a leading role in the development, and it should be so committed to the task that it should probably have carried out the development on its own, even if the ECE/METIS group had not existed;

- the statistical office with the leading role should (provided that the development as such is successful) be prepared to take the responsibility for the maintenance of the software during a reasonable time period.

3.3 Architecture and features of a pilot metainformation system

3.3.1 Some lessons from metainformation failures in the past

Many metainformation systems in the past have failed, among other places also in statistical offices. Some of the most critical reasons for these failures are:

(3.3) Metadata collection is (like most data collection activities) dull, expensive, and time-consuming;

(3.4) The "natural" suppliers of metadata, those who know something about the object data to which the metadata refer, are difficult to motivate, since, by definition, they do not themselves need the metadata that they must supply to the metainformation system; at least they do not need the metadata as a formalized part of a system - they have the metainformation in their own heads;

(3.5) The "natural" consumers of the metadata, on the other hand, will not find a metainformation system to be of very much value, until the metadatabase is reasonably complete, that is, until it covers most of the object data to be covered; this is so, because users of data (and not least users of statistical data) are usually interested in several (related) collections of data, which have been collected and described by different people and different organizational units;

(3.6) Partly as an implication of (3.4) and (3.5) there is an unfortunate disjunction between the "natural" users and producers of metadata: those who need the metadata cannot themselves supply the metadata, and even has very little control, if any, over the supply side, whereas those who must supply the metadata, because they "own" the knowledge about the object data, do not benefit very much from the formalized and automated availability of the metadata;

(3.7) If we compare a metainformation system with a well-functioning market, we will find that a typical metainformation system from the past has been lacking some important features:

- relations between the supply side and the demand side have been virtually non-existing;

- the supply side has had most of the costs but little benefit, whereas the demand side, which has been intended to get the benefits, have not had to pay for them, or, if they have had to pay (maybe as an integral part of the whole information service), the payments have not been routed directly to the metadata suppliers, and have thus failed to have the "self-regulating" effect that such payments usually have in a market.

3.3.2 Implications for future metainformation systems

From the past experiences we can learn some lessons concerning the desirable architecture of future metainformation systems:

(3.8) Metadata collection activities should be minimized in the sense that no piece of metainformation should have to be entered into any metainformation system more than once, and derivable metadata should be automatically derived rather than manually entered;

(3.9) Huge retrospective metadata collection activities should be avoided; instead as much as possible of the metadata input flow should be generated as a side-effect of other activities; for example, the more or less formalized models and descriptions that are typically generated by systems analysis and design activities should be automatically captured and organized as potential metadata for the information system under development;

(3.10) Some type of formalized cost/benefit mechanism needs to be introduced into the architecture of a metainformation system in order to relate users and producers of metadata in a healthy and constructive way; the mechanism needs to be relatively sophisticated, since there is a many-to-many relation between users and producers of (meta)data: the same (meta)data may be used by many different users, and the same user may use (meta)data concerning many different but related data collections from many different producers.

3.3 Some implications for statistical metainformation systems

The statistical information systems, often called (statistical) surveys, organized and operated by statistical offices and other statistical organizations, are valuable assets of such organizations. However, without properly integrated metainformation (sub)systems, the value of the information systems as such is drastically reduced. Since today's statistical information systems are by and large formalized and computerized, the properly integrated metainformation systems must also be formalized and automated, if the pace of the metainformation flow is to keep up with the pace of the information flow.

It may be true that there are some exceptional statistical surveys, which are self-contained in the sense that they have a well-defined set of users, who essentially need data from this survey alone, and who can rely on person-to-person contacts with the producers of the survey, together with oral traditions between themselves,

for satisfying their needs of metainformation. However, most users of statistical information need to be able to interpret and combine data from many different sources, and different users and usages will need different combinations and analyses to be made.

4 Outline of an architecture for a statistical metainformation system

An important principle, know from some methodologies concerning information systems design, is that the structure of the information system should, as far as possible, reflect the structure of the object system of the information system. We shall apply this principle to the subsequent analysis of a suitable structure and architecture for statistical metainformation system. Since the object system of a statistical metainformation system is a statistical system, or a statistical survey, we should start with an analysis of the structure of a statistical survey.

4.1 The gross structure of a statistical survey

The kernel of a computerized statistical survey is a process, or subsystem, which transforms certain object data, with accompanying metadata, into other object data and metadata:

(4.1) {object data} + {metadata} → [survey processing] → {object data} + {metadata}

(In a formula, like (4.1), {...} indicates a set of information or data, and [...] indicates a process that processes information or data.)

The formula (4.1) can also be illustrated by a flow diagram, as in figure 4.1, where a shaded box indicates an information/data set, and an unshaded box indicates a process. A flow diagram should be read top-down and from left to right, if nothing else is indicated by arrows.

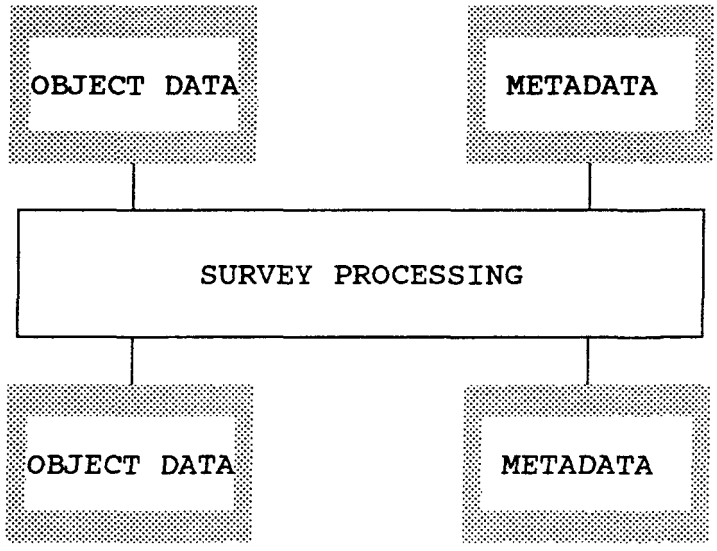

Figure 4.1. *The gross structure of a statistical survey.*

The input object data of a statistical survey are typically microdata, the survey processing typically includes one or more aggregation processes, and the output data are typically macrodata.

The metadata accompanying the input object data are, in today's practice, usually far from complete. They do not satisfy all the potential purposes of metadata as described in chapter 2 of this paper. The computerized subset of the input metadata will often be limited to those aspects of the input object data, for which the software products used in the survey processing will require computerized metadata, such as file and record descriptions and some textual data to be used as output metadata in the presentation of the output object data.

4.2 The object data processing of a statistical survey

In the following we shall describe the typical steps in the processing of the object data in a statistical survey a little more in detail. The discussion is illustrated by figure 4.2. (In section 4.5 we shall return to the processing of metadata.)

The input object data to the survey processing come from a data collection process or subsystem. There are several types of data collection that produce the input object data of a statistical survey:

- **Direct observations and measurements.** This is the most typical form of data collection in a "classical" statistical survey. **Variables of interest** are directly observed or measured for (in the case of a **total survey**) *all*, or (in the case of a **sample survey**) *some* **objects of interest**, belonging to one or more **object types**, and to one or more object collectives or **populations of interest.** The observations/measurements can be made by **interviewers**, or by the objects themselves (if they are persons), or by persons who are somehow related to the objects of interest (if the latter are not persons). **Questionnaires** in some form or other are often used as measurement instruments.

  Data emanating from direct observations and measurements will be called **direct data** or **direct observations.**

- **Indirect observations and measurements.** This form of data collection is of growing importance for statistical offices and organizations. Indirect observations and measurements are observations and measurements that were originally made

  - *either* by some other statistical survey, inside or outside the statistical office under consideration,

  - *or* within some non-statistical information system, for example an administrative information system operated by some administrative agency.

  Data emanating from indirect observations and measurements will be called **indirect data** or **indirect observations.**

- **Derived data collection.** Derived data are data, which have not been directly observed or measured, neither within the survey under consideraion, nor within some other information system; instead they have been derived from such data by means of some procedure, computerized or other.

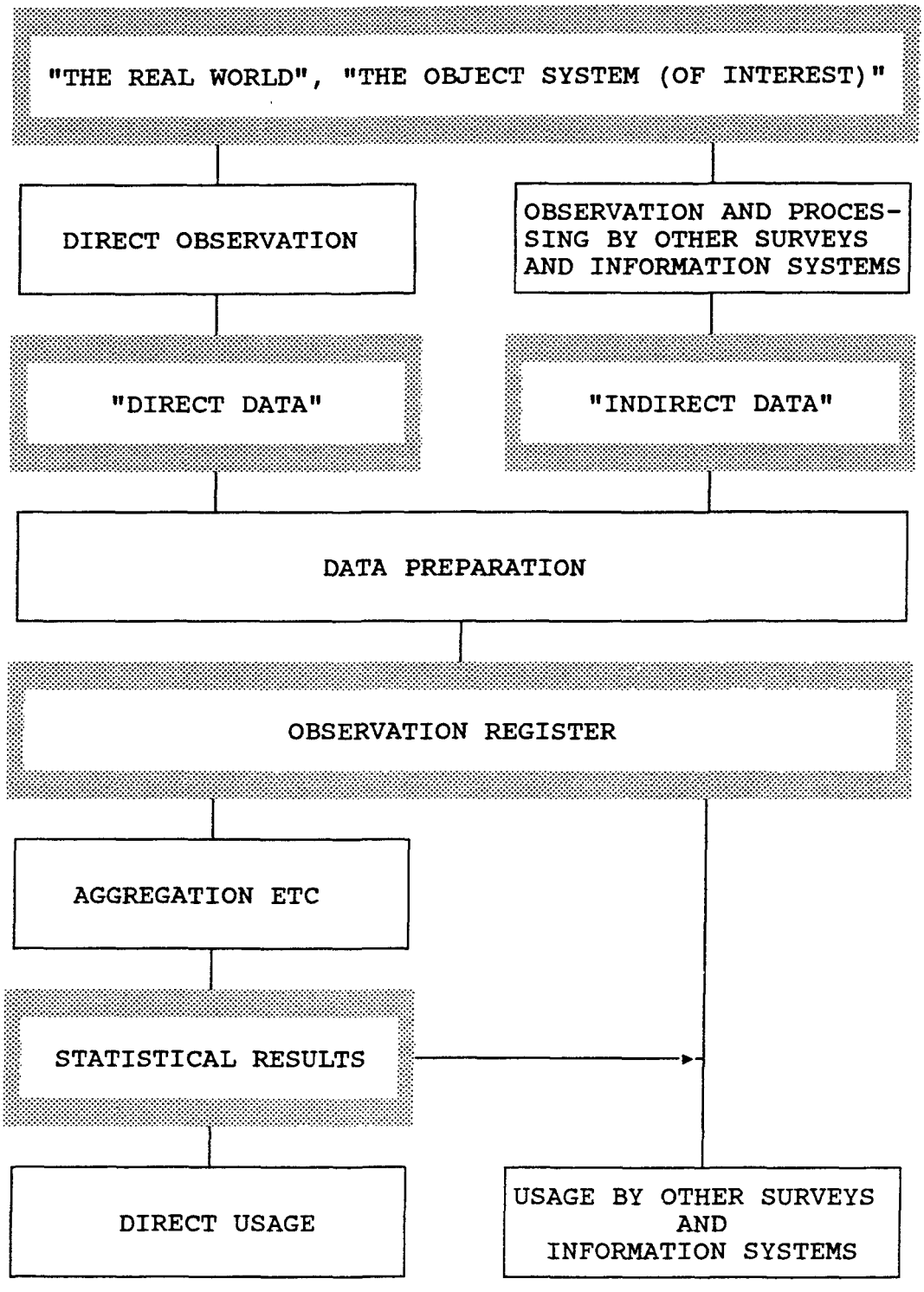

Figure 4.2. *The object data processing in a statistical survey.*

There is no sharp borderline between, on the one hand, direct/indirect observations and measurements, and, on the other hand, derived data. "Raw" observations and measurements are very seldom used in survey processing "just as they are"; they are usually checked and edited in various ways before the are actually processed - however, in this context we shall not usually regard such procedures as editing as "derivation procedures". Instead, we shall use the term **data preparation** to cover all processing (including derivation) of the data collected by a survey (directly or indirectly) that occurs *within* the survey and *before* a **(final) observation register** (see below) is established. The data preparation may include such operations as editing and coding, as well as some basic computations of derivable variables. Data preparation and derivation that indirect observations have possibly been subject to in the other surveys and information systems that they emanate from, will be regarded as **external** to the survey under consideration. (This does not exclude the possibility that such data will be subject to further data preparation as **internal** operations within the survey under consideration.)

At some stage in the execution of a statistical survey we will (at least temporarily) "close" the data collection and data preparation activities and establish a **(final) observation register**. A final observation register may be physically organized as a set of several computerized files.

An observation register has two important functions in a statistical information system. One is to serve as the basis for **aggregation** and **estimation** procedures, producing the (immediate) **statistical results** of the survey. The observation register could thus be regarded as the natural basis for the regular, "first-time" statistics production, for which the survey is responsible.

The other important function of an observation register is to store some version of the input object data of a statistical survey for **potential future (re)use** by researchers and others. The observation register can be stored for this purpose in some **data archive** and/or in more "active", computerized **databases**. In the past, the orginal files of manually filled in questionnaires were often a principal component of the observation register, but today such manually readable documents are not regarded as very useful for future users and are usually replaced by computerized (or at least microfilmed) versions.

Aggregation can be regarded as a special case of data derivation, whereby macrodata are computed from microdata. Thus the gross structure of a statistical survey includes the possibility that some (or all) of the object data input to a statistical survey consists of macrodata aggregated by another survey. Defining a statistical survey in this way, we include in the survey concept such information systems, like the *System of National Accounts* (SNA), which are more or less entirely based on macrodata, which have been aggregated by other surveys.

The statistical results of a survey will often be used directly by some client users of the statistical information system. Various forms of **statistical reports and publications** are the traditionally typical **channels** between the statistics producer and the client users. Today new channels, like **statistical databases**, are becoming increasingly popular, especially as a basis for downloading statistical results from the computers of the statistics producer to the computers (and software) of the clients.

The term "**statistical database**" is often used to denote *only macrodatabases* with (more or less) final statistical results, as just described. However, we shall use the term in such a way that *also microdatabases* like the **observation register** mentioned above, could be included in the concept of a statistical database, even if it has to be treated with greater care from security point of view.

Thus, what we may call the **statistical database of a statistical survey** has two major components:

(4.2) an **observation register**, typically containing microdata; and

(4.3) a database of **statistical results**, typically containing macrodata.

The purpose of the statistical database and its two major components is also twofold: *immediate* and *mediate* production and usage of statistical information. For example, an observation register serves as an **immediate** basis for production of statistics, whenever tables and analyses are produced directly within the same survey that created the observation register. On the other hand, whenever an observation register, after the register-creating survey has been completed, serves researchers and others carrying out special studies, the observation register is a **mediate** basis for production of statistics.

Similarly, the statistical results stored in the statistical database of the survey could have an *immediate* as well as a *mediate* role in the usage of statistics; the latter is the case, whenever the statistical results are used by other surveys than the one originally producing and storing the statistical results.

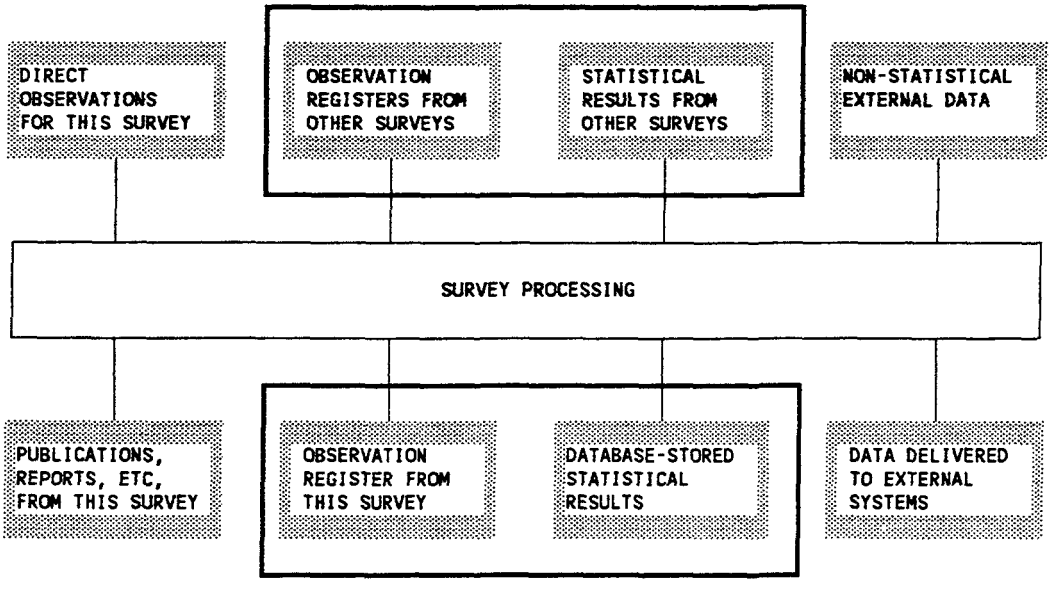

Figure 4.3. *The inputs and outputs of a statistical survey.*

Figure 4.3 summarizes what we have said so far about the input to, and the output from, a survey process. The fat line on the output side embraces the **statistical database** of the survey under consideration. The corresponding fat line on the input side embraces (a subset of) the **consolidated statistical database of related surveys.**

4.3 Preparation of data collection

In the previous section we discussed the different steps in the processing of object data in a statistical survey. However, before we can even start the data collection, we must prepare the survey by performing such tasks as

- establishing a **frame**, that is a list of **elements**, which are in a well-defined way related to

  - data sources for the survey;
  - respondents, contact persons;
  - objects of observation;
  - objects of interest;

  (some of these entities may be overlapping);

- drawing a **sample**, usually **random**, from the frame, thus establishing (through the **frame procedure**) which contact persons should be approached (and how to approach them), which objects to be observed, and which objects of interest to derive values of variables for;

- establishing an **"empty" observation register**, that is, a "skeleton" of one or more matrixes, containing a row for every object to be observed (directly or indirectly) and a column for every variable to be measured.

We shall use the term **survey preparation** to denote the survey operation covering this type of tasks.

Survey preparation should not be mixed up with **survey planning** (cf section 4.4). The tasks involved in survey preparation are the first operative tasks that are executed once the survey plan has been established. Design decisions concerning object types and populations of interest, variables of interest, object types and variables to be observed, frame and frame procedure, sampling strategy, etc, belong to survey planning, and not to survey preparation.

Figure 4.4 illustrates the principal operation steps and information structures involved in the operation of a statistical survey.

Figure 4.5 shows the structure of survey operation by means of a so-called JSP-diagram (JSP = Jackson Structured Programming). The figure should be "read" as follows:

"Survey operation consists of *first* survey preparation, *and then* data collection, *and then* observation processing, *and finally* output preparation."

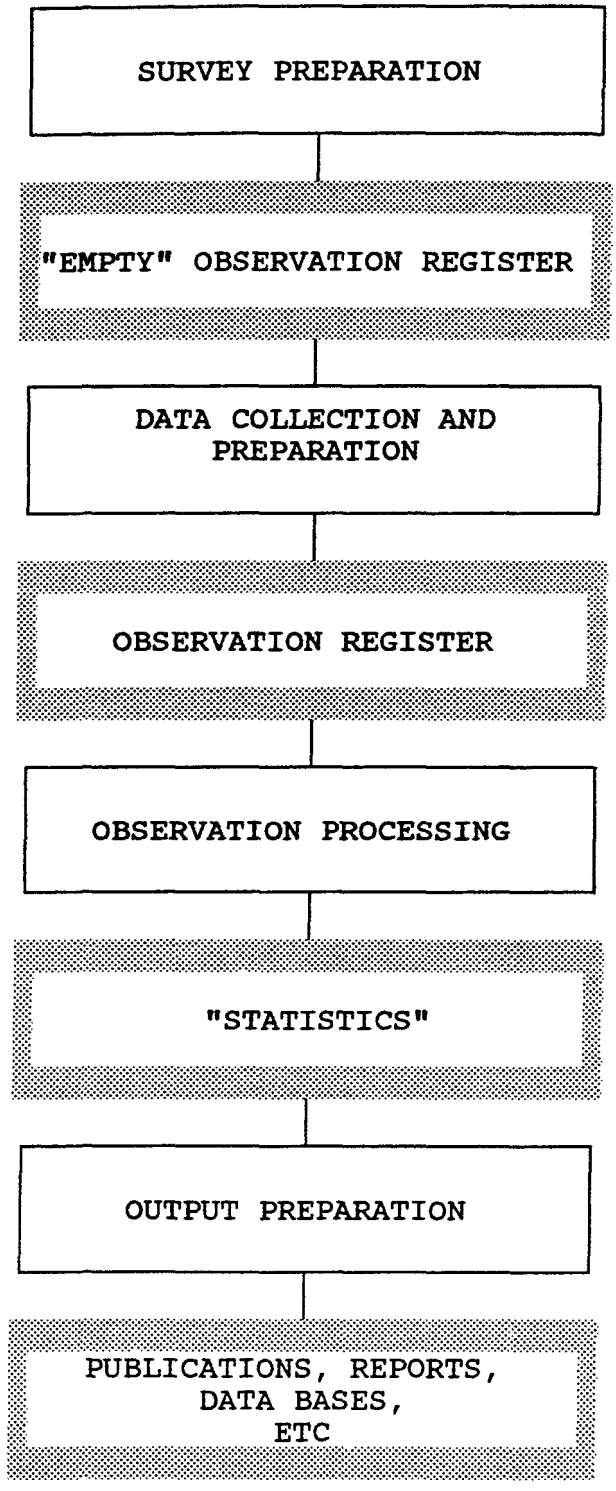

Figure 4.4. *The principal steps in the operation of a statistical survey.*

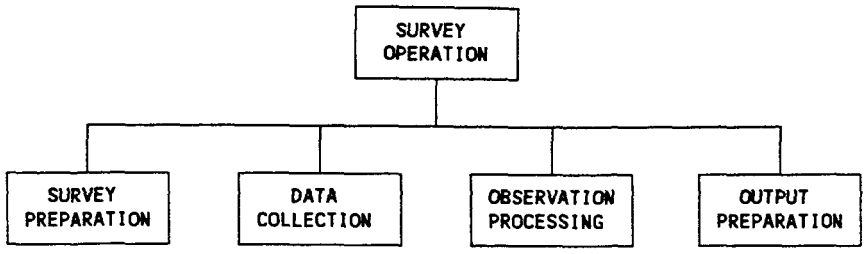

Figure 4.5. *The structure of survey operation illustrated by a structure diagram of JSP-type.*

4.4 Survey control and the survey life cycle

Figure 4.6 illustrates the whole **survey life cycle** by means of a JSP-diagram. The life cycle consists of three major phases: survey planning, survey operation, and survey evaluation. So far we have only talked about the survey operation phase and the steps that this phase is made up of. However, especially in the analysis of statistical metainformation and metainformation systems, it is essential to consider also the planning and evaluation phases. From a systems point of view, planning and evaluation are **control systems** visavi the survey operation system.

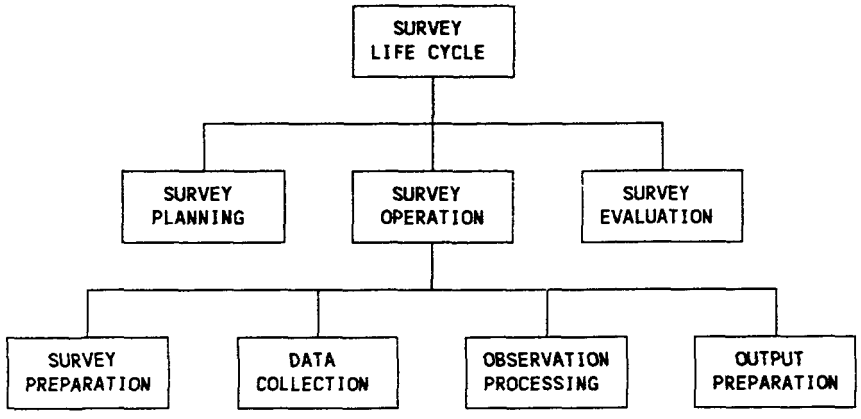

Figure 4.6. *The survey life cycle.*

4.5 The metadata processing of a statistical survey

Now we can turn our interest to the metadata parts of the formula

(4.1) {object data} + {metadata} → [survey processing] → {object data} + {metadata}

(Cf also figure 4.1 above.)

A complete set of input metadata should contain the metadata necessary for

(4.4) performing all survey processing tasks, from data collection and data preparation to the production of the output object data on the basis of the contents of the observation register of the survey;

(4.5) correct (future) interpretation of the data in the observation register of the survey, when (future) users are (re)using these data;

(4.6) producing the output metadata describing the output object data.

There are two major sources of survey metadata: **survey design decisions** and the **survey process** itself.

Survey design decisions determine models and concepts underlying the survey:

- what real world problems was the survey intended to highlight?
- what decisions was the survey aimed to support?
- what domain of interest should the survey help to understand?
- what questions should the survey try to answer?

Metadata emanating from survey design decisions typically have the form of more or less formalized model descriptions and concept definitions. Most of these metadata can be captured *before* the survey is actually carried out in terms of object data collection, preparation, and processing.

The survey process itself generates metadata about what actually happened *during* the survey, for example,

- what problems occurred during observation and measurement?
- how did the respondents react to different questions in the questionnaire?
- what were the sizes of non-response for different object types and different variables?
- what was done to prevent non-response, and what was done to compensate for it when it had occurred?
- how were the input data checked, edited, and/or corrected?

The survey design decisions are typically taken during the **survey planning** phase of the survey life cycle (cf figure 4.6 above). Thus the survey planning phase generates most of the metadata concerning models and concepts used by the survey. Similarly the **survey operation** phase of the survey life cycle generates most of the metadata about what actually happened during the survey. Finally, the **survey evaluation** phase generates more "in depth" metadata and analyses concerning various aspects of the quality of the survey and the data emanating from the survey.

We shall now use an **input/output/process scheme of analysis** in order to find out more details about the metadata (and object data) processing that takes place during each one of the three major phases of the survey life cycle.

4.5.1 Survey planning

Input: survey requirements;

local metadata, that is, metadata which are local to the (series of) survey(s) under consideration, including:

- plans and specifications used by earlier survey repetitions,
- experiences from earlier operations of the survey (series);

global metadata, that is, metadata concerning surveys related to the survey under consideration, including:

- other survey plans and specifications,
- experiences from operations of other surveys;
- metadata concerning common resources, such as
    - statistical standards;

Output: survey plan containing:

- specification of domain of interest (populations, variables),
- specification of planned survey output information,
- specification of planned survey input procedures:
    - frame procedure,
    - sampling procedure (if relevant),
    - data collection procedure:
        - sources of information,
        - contact procedures,
        - measurement procedures,
        - measurement instruments,
    - observation register,
    - data preparation procedures:
        - data entry,
        - coding,
        - editing,
- specification of planned survey processing procedures:
    - observation models (e g non-response),
    - estimation models,
- specification of survey output procedures;

specification of data processing system containing:

- specification of survey preparation subsystems,
- specification of survey input subsystems,
- specification of survey processing subsystems,
- specification of survey output subsystems;

updated local and global metadatabases;

Process: requirement analysis, statistical design, systems design;

4.5.2 Survey operation

Input: object data obtained by direct observation or from the survey database;

survey system knowledge base containing:

- metadata about object data in the survey data base,
- plans and specifications for this and other surveys,
- metadata about previous operations (repetitions) of the survey under consideration;

data (object + meta) from other surveys and information systems;

Output: presented results and analyses from this survey, including:
- object data,
- metadata;

object data in the survey database, updated with:

- observation register from this survey,
- statistical results (macrodata) from this survey;

survey metadatabase, updated with:

- metadata about new observation register,
- metadata about new statistical results,
- revised survey plans and specifications,
- metadata about this survey operation;

data (object + meta) communicated to other surveys and information systems;

Process: survey preparation,
data collection and data preparation,
estimation and other computations,
output processes;

4.5.3 Survey evaluation

Input: updated version of the survey database, including:

- observation register produced by the survey,

- statistical results produced by the survey;

updated version of the survey metadatabase, including:

- survey plan,
- metadata about the survey operation,
- feed-back from the users of survey data;

Output: evaluation reports,
updated metadatabases,
(improvement) proposals for future surveys;

Process: evaluating analyses,
evaluation surveys,
metadata processing;

4.6 A statistical information system as a system of related surveys

So far in this analysis we have looked upon the structure of a statistical survey from the perspective of *one single statistical survey*. This may be a good starting-point, but it is certainly not enough. We have already caught glimpses of *"related" statistical surveys*, that is, statistical surveys that are related to the one under consideration, either because it produces some input to the survey under consideration, or because it uses some output from the survey under consideration (cf figures 4.2 and 4.3 above).

However, we may go much further and claim that the very "reason for being", or *raison d'être*, for statistical offices and other statistical organizations is that individual statistical surveys are related to each other in rather complex patterns, forming **statistical information systems**, or simply (for short) **statistical systems.**

Depending on circumstances the statistical system under consideration may be a very large and complex system, like the statistical system for which a certain statistical office is responsible, or the statistical system of a whole country or even a whole world community of countries. Equally often, however, the statistical system under consideration may be "only" some subsystem of such an extensive system, for example the System of National Accounts of a country, a health-statistical system, or a socio-demographical system. Figures 4.7 and 4.8 indicate a structure and a terminology, which can be used for all types of statistical systems, large or small, complex or less complex.

In figures 4.7 and 4.8 the term **survey universe** is used to denote the particular statistical system under consideration, regardless of its size and complexity. The survey universe may consist of subsystems, here called **survey families.** A survey family may again consist of subsystems, which are also survey families. Finally one will reach the level of the **individual surveys**, or **survey members**, of the survey families. Individual surveys, which are repeated in more or less the same form, at more or less regular intervals (monthly, quarterly, yearly etc) are said to be **survey rounds** or **survey repetitions** within one and the same **survey series.** In figure 4.7 survey series and survey rounds are explicitly recognized, whereas in figure 4.8 a survey series is regarded as a special case of a survey family.

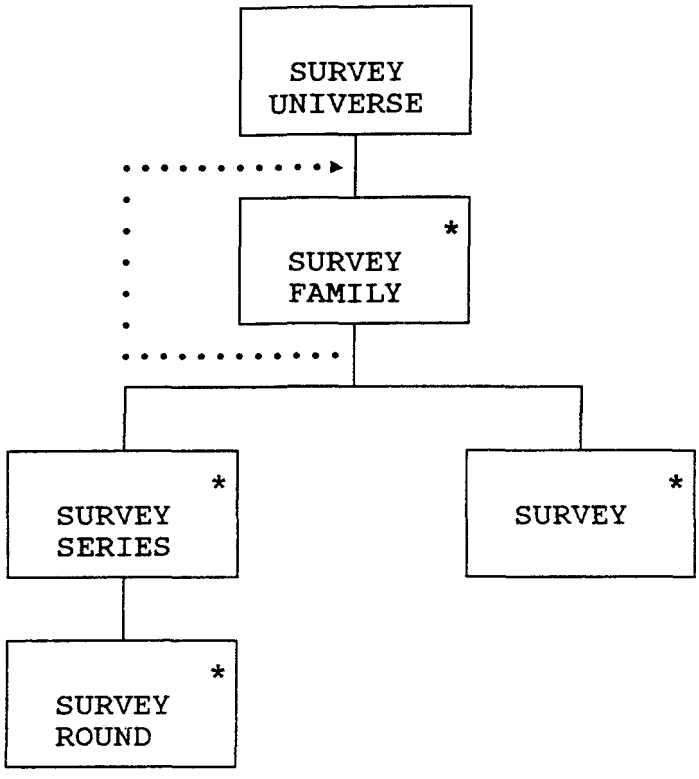

Figure 4.7. *A statistical system, or survey universe, consisting of (possibly several levels of) survey families, where each family will consist of some survey series and some stand-alone surveys; a survey series consists of surveys or survey rounds.*

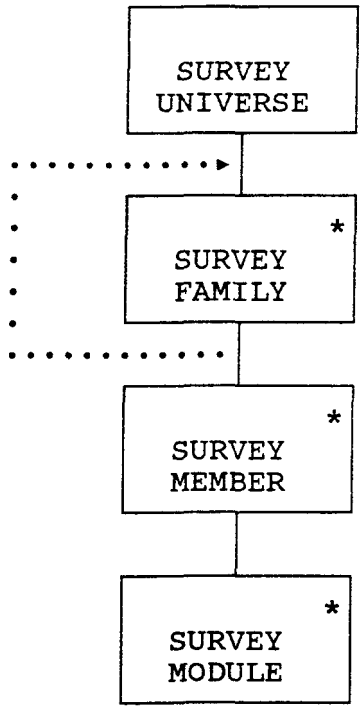

Figure 4.8. *An alternative version of figure 4.7.*

An individual survey (survey member) may in turn consist of "subsurveys" or **survey modules**. For example, a consumer price index (survey) may be made up by several, more or less independent subindexes (survey modules).

A note on survey series

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 variable 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.

End of note on survey series

Earlier in this paper we have discussed operation flows and control flows for individual surveys (survey family members). Similar discussions may be carried out for collections of related surveys, like survey series, survey families, and survey universes. For example, figure 4.9 shows the life cycle of a survey family, consisting of a planning phase, an operation phase, and an evaluation phase.

On the structural level figure 4.9 looks very similar to the life cycle of an individual survey, as visualized in (the upper part of) figure 4.6. However, there are some important differences. For an individual survey the three major phases are normally executed in a relatively serial way: *first* the survey is planned, *then* it is operated, *and finally* it is evaluated. For a collection of related surveys, like a survey family or a survey universe, the three types of activities (planning, operation, and evaluation) are **interleaved** (rather than serial) in the sense that they are going on more or less in parallel: some surveys (or parts or aspects of surveys) in a collection of related survey may be operated or even evaluated at the same time as some other surveys (or parts or aspects of surveys) are still only in the planning phase. Furthermore, there are surveys of a more or less continuously ongoing nature, like **registers** and **event-based statistical systems**.

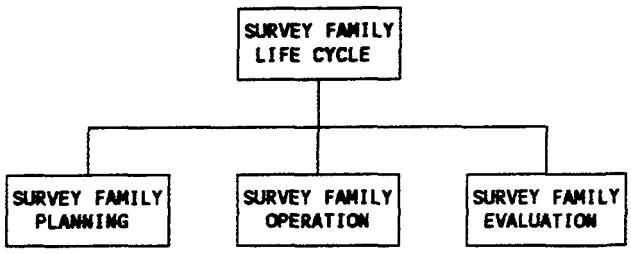

Figure 4.9. *The life cycle of a family of related survey members.*

4.7 Survey (series) knowledge bases and survey (series) expert systems

For series of "similar" surveys of the type discussed in the previous section, statistical offices usually create relatively stable organizational units, which are given the responsibility for operating and maintaining the survey series. Over time such an organization normally accumulates considerable knowledge about the survey series, or the **statistical product**, as the survey series is sometimes called. Parts of this knowledge may be documented in manuals, handbooks, work instructions, system documentations, publications, methodology reports, etc, but substantial parts of the knowledge is often **tacit knowledge** in the sense that it is not at all documented, but only "known" (consciously or unconsciously) by (some of) the people in the organization.

There are many good reasons for organizing the knowledge about a statistical product in some sort of **computerized knowledge base**, including existing formal and informal documentations, as well as tacit knowledge, obtainable only from the people involved in the work with the statistical product. Some of the arguments for such a **survey (series) knowledge base** are:

(4.7) The survey series as such (and certainly the data from the survey) will normally have a longer life time than any staff member working with the survey;

(4.8) New staff members have to be trained, and a well organized knowledge base, containing (meta)information about the survey, its data, and how it is carried out, will make the training more efficient;

(4.9) Even if a user of the statistical product can obtain relevant metainformation by means of personal contacts with the staff responsible for the product, this "retrieval mode" may sometimes be very inefficient, especially if the user is interested in statistical data (and metadata) from several statistical products;

(4.10) Computerized survey knowledge bases would, if properly integrated with other components in a system of statistical information systems, be able to serve as a basis for automated production of other essential metadata in such a system, for example, metadata to accompany archival data, metadata required by the software used for processing the object data in the survey, and metadata needed for quality declarations of the output from statistical surveys.

On a higher ambition level, a survey (series) knowledge base could be used as a component of a **survey (series) expert system**, an instrument, which could be used in the training of staff members and users, or as a "computerized expert" for relatively routine consultations concerning the statistical products.

4.8 Survey-independent metadata

It seems appropriate to start the specification of metadata items (metadata message types) for a system of statistical information systems, as we have done here, *from the individual surveys*. Many metadata items *are* **survey-specific**. At the same time one should also keep in mind that there are also many metadata items which are **survey-independent** in the sense that the "natural" object part of the metadata message type is not a survey, but some other information system component, for example, a variable, an object type, or a population. Such metadata are usually common to a smaller or larger number of surveys, forming what we have called above a **collection of related surveys** (for example, a survey family, or a survey universe). Survey-independent metadata are **global metadata** with respect to the individual surveys.

Statistical standards (standard definitions, standard classifications, etc) are good examples of metadata, which have a high degree of survey-independence in the sense that they are common to a large number of surveys. The degree of survey-independence is higher, the more universally established the standard is. For example, national standards are more survey-independent than office-internal standards, international standards are more survey-independent than national standards, etc.

The left part of figure 4.10 shows a simplified version of the structures in figures 4.7 and 4.8. The right part of figure 4.10 indicates the databases, or **information bases** (object + meta), with which each level of survey collection typically interacts, during each of its life cycle phases.

During the planning, operation, and evaluation of a survey universe, there are interactions between survey universe level processes and survey universe data and metadata; these data and metadata are global with respect to survey families and survey family members.

During the planning, operation, and evaluation of a survey family, there are interactions between survey family level processes and survey family data and metadata; these data and metadata are local with respect to the survey family, and global with respect to the individual surveys in the survey family.

Finally, during the planning, operation, and evaluation of a survey family member, there are interactions between the individual survey level processes and survey data and metadata; these data and metadata are local with respect to the individual surveys.

The structure visualized in figure 4.10 represents but *one* way of organizing the object data and metadata of a universe of statistical surveys, for example the universe of statistical surveys managed by a statistical office. Naturally, one could imagine alternative structurings and organizations. However, the one presented here seems to be in harmony with contemporary ambitions to organize the activities of

statistical organizations with an optimal trade-off between, on the one hand, decentralized control and responsibility and, on the other hand, coordination of related branches of statistics.

From a more database-technical point of view, the proposed organization seems to be based on the principles of non-redundancy and normalization known from conceptual data modelling and relational database theory.

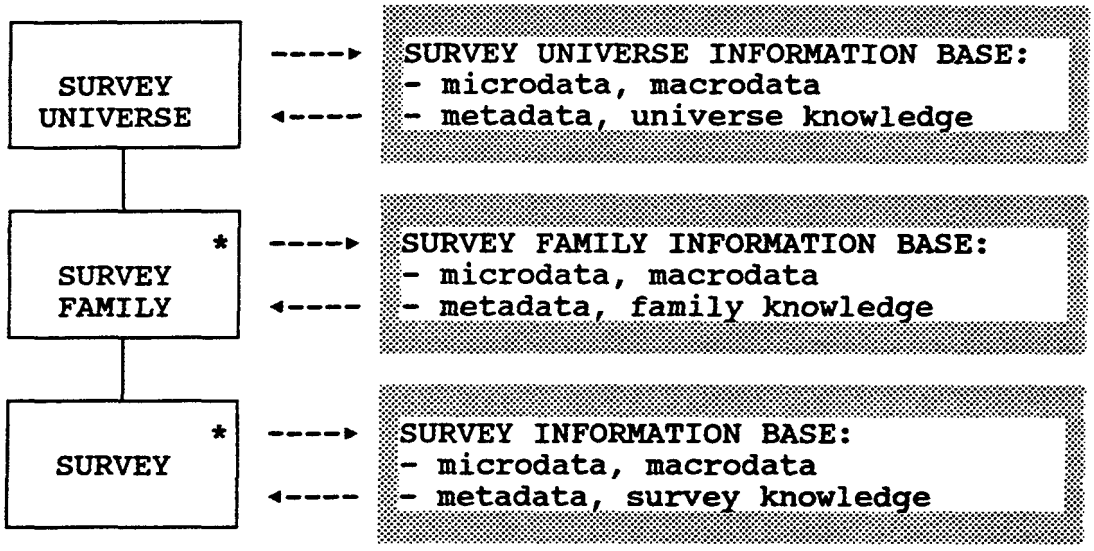

Figure 4.10. *A proposed structure for the processes and databases of a statistical system.*

5 Future work

Hopefully this report has contributed some ideas about how the **bases** as well as the **flows** of well integrated **object data** and **metadata** of a universe of related statistical surveys could be best organized and managed. I think that it is not an exaggeration to say that it is a question of survival for statistical organizations to solve these problems better in the future than we have done in the past. Without an adequate and efficient handling of metadata (as well as of object data) statistical organizations will not be able to offer the services expected from them in an era of advanced information technology. It is not enough any longer to tackle these problems survey by survey only. A systems approach must be applied, and this report has proposed such an approach. It will be an important task for the UN/ECE METIS Group to investigate this and other approaches, taking into account both conceptual aspects and implementation considerations. (In the view of this author, the former should in principle precede the latter, but of course it should not be "forbidden" to do some experimentation with implementation, even before a "perfect" conceptual design has been worked out.)

On the basis of the framework presented in this report, I would propose the following list of important tasks for the METIS Group. It should go without saying that the list is not intended to be complete.

5.1 Analyze metainformation needs for a statistical system

According to this report, for a statistical system to be **infologically complete**, it should contain the metadata which are necessary for proper interpretation of the output data from the system. The output data consists of two major parts:

- statistical results (macrodata); and
- observation registers (microdata).

As a starting point for the analysis and specification of these metadata needs.

The metadata necessary for proper interpretation of observation registers have been analyzed in reference (6.1) by Bengt Rosén (professor of Statistics) and Bo Sundgren.

The metadata necessary for proper interpretation of statistical results have been analyzed in reference (6.2) by Bo Sundgren.

In a sense, **infological completeness**, as defined in this report, is a minimum requirement on the metadata for a statistical system. Without the metadata necessary for interpreting the object data, it would not be meaningful to produce and store the object data. However, as is indicated in chapter 2 of this report, there are other metadata needs for statistical surveys and statistical systems, which go beyound the requirements of infological completeness. For example, a well designed metadata system may be very helpful in locating and accessing object data of potential interest to a client user, who may perhaps only be able to specify his or her problems and information needs in a relatively "fuzzy" way. Even if the user is able to give a more formal specification of information needs, it may not be a trivial task to determine if and where relevant data are available within a statistical system, especially if the statistical system consists of many individual surveys, and

the user cannot specify the particular survey(s), where the data might be available.

Procedural completeness is another requirement on metadata for statistical surveys and statistical systems, which has been mentioned in this report, and which would be interesting to analyze further. The concept leads naturally to ideas concerning **expert systems** and other **knowledge-based systems**, supporting different operations and phases in the life-cycle of statistical surveys and statistical systems.

The analysis of metainformation needs should cover both **factual metainformation** and **rule-based metainformation.**

5.2 Find "natural" metadata sources and "convenient" collection procedures

When we have identified and defined certain metadata (output) needs, the next question is to identify and define the (input) metadata, which are necessary for the production of the desired output metadata, and, not least important, to determine how, and from which sources, the metadata should be collected.

Given an appropriate conceptual framework, like the one outlined in this report, it is more or less a purely logical task to infer which input metadata are necessary for the production of certain output metadata. This process of logical inference is sometimes called **precedence analysis** in information systems theory, since it amounts to finding the information precedents of certain specified information kinds (information succedents).

However, the second part of the problem, that of finding "natural" metadata sources and "convenient" metadata collection procedures, requires not only logical proficiency, but also good organization skills. As was indicated in section 3.3.1 of this report, many metadata projects in the past have failed partially because they have failed to design metadata collection procedures, which do not require too much effort from people in the organization, who are not very highly motivated for producing documentations and other forms of metadata, at least not if this has to be done as a separate activity, which is not very well integrated with the "main flow" of operational work.

A "convenient" metadata collection procedure could be defined as one, where the metadata are not at all collected as a separate activity; rather the collection of metadata should be an automatical side-effect of doing something else, which *everyone involved recognizes as something that has to be done under all* circumstances, ideally for very operational reasons.

5.3 Establish a reference conceptual model for statistical metainformation

It should be a task of the UN/ECE METIS Group to propose a conceptual model for statistical metainformation, which could serve as a standard or reference model for statistical organizations. The model should be worked out in accordance with some formalism for conceptual models, for example the Object-Property-Relation (or Entity-Attribute-Relation) model. Emerging ideas from Object-Oriented analysis and design of information systems could also be considered in this connection.

The concepts and models lined out in this report could serve as a starting-point for the development of a reference conceptual model for statistical metainformation.

However, a lot of detailed work remains to be done. The model should be tested (for consistency and otherwize) against the results of the analyses indicated in 5.1 and 5.2 above, which could be carried out in parallel with the work on the reference model.

5.4 Establish a reference flow model for statistical metainformation

We have already several times noted how important it is for the metadata flows of a statistical system to be very well integrated with the object data flows and work procedures. It could in fact be worthwhile to formalize not only a conceptual model for statistical metadata, but also a **flow model**, describing the flows of metadata in a typical statistical system, starting from "natural" metadata sources (cf 5.2) and ending with desirable metadata outputs (cf 5.1).

It would also be worthwhile to think about optimal implementations of flow models for statistical systems. One criterion for the optimization should be that metadata (like object data) should be captured (and translated into computerized form) only once, and with as little manual effort as possible. Another (consequence) criterion should be that metadata should be transformed automatically (to the maximum extent, which is logically possible) together with the object data that they describe. An algebra for metadata transformations could maybe be established, in analogy with object data algebras like the relational algebra, known from relational database theory, or the base operator algebra, known from the UN/ECE SCP projects. Ideally, object data algebras and metadata algebras should be integrated.

5.5 Establish *some* important metadata interfaces for statistical systems

It may be difficult to establish complete metadata reference models (conceptual model + flow model) of sufficient generality, precision, and, not least important, acceptance by statistical organizations. Rather than standardizing complete reference models, one could focus on certain particularly important **interfaces** within such models. This would be in line with, and should be coordinated with, other international standardization work that is going on at present, notably the UN/EDIFACT standardization efforts.

5.6 Make systematical studies of ongoing statistical metainformation projects

There seems to a "new wave" of rather ambitious metainformation-related project sweeping over statistical organizations at present. This author knows about such projects at the Australian Bureau of Statistics, Statistics Canada, Statistics Sweden, and in the Organization of Economic Cooperation and Development, and there are probably several others. It seems to be a very relevant and natural task for the UN/ECE METIS Group to make studies of these projects, and to test its own ideas against the ideas of the other projects.

5.7 Experimental design and implementation of software and other tools

In the opinion of this author, development of software and tools does not lend itself very easily to international cooperation. However, this should not exclude that some experimental development of this nature could be undertaken by the METIS group. In addition to software components, certain types of database components (for example, catalogues of standards) could be considered.

6 References

(6.1) Bengt Rosén & Bo Sundgren: *"Documentation for reusage of microdata from the surveys of Statistics Sweden"*. Statistics Sweden 1991. Bengt Rosén is a professor of statistics at Statistics Sweden. Unfortunately this report is at present only available in Swedish; the Swedish title is: *"Dokumentation för återanvändning av mikromaterial från SCBs undersökningar"*.

(6.2) Bo Sundgren: *"What metainformation should accompany statistical macrodata?"*. Report written 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.

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 1991:

1991:1 (grön)	Computing elementary aggregates in the Swedish consumer price index **(Jörgen Dalén)**
1991:2 (grön)	Översikt av estimatorer för stora och små redovisningsgrupper **(Sixten Lundström)**
1991:3 (grön)	Interacting nonresponse and response errors **(Håkan L Lindström)**
1991:4 (grön)	Effects of nonresponse on survey estimates in the analysis of competing exponential risks **(Ingrid Lyberg)**
1991:5 (grön)	Study of the estimation precision in the Chinese MIS surveys **(Bengt Rosén)**
1991:6 (beige)	Abstracts I - Sammanfattning av metodrapporter från SCB
1991:7 (grön)	Kvalitetsfonden 1990: Projekt - handläggning - utvärdering **(Roland Friberg och Håkan L Lindström)**
1991:8 (grön)	Tre återintervjustudier: Inkomstfördelningsundersökningen **(Jan Eriksson)**, Högskoleenkätuppföljningen **(Anders Karlsson)**, Årsarbetskraften **(Majvor Karlsson och Solveig Thudin)**
1991:9 (gul)	What metainformation should accompany statistical macrodata **(Bo Sundgren)**
1991:10 (grön)	The Family Expenditure Survey: An Experiment with Incentives **(Håkan L. Lindström)**, Seasonal Variation and Response Behaviour in Swedish Households Expenditure **(Peter Lundquist)**, Reducing Nonresponse Rates in Family Expenditure Surveys by Forming Ad Hoc Task Forces **(Lars Lyberg)**, A Study of Errors in Swedish Consumption Data **(Martin G. Ribe)**

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/ADB, SCB, 115 81 STOCKHOLM, eller per telefon 08-783 41 47.

SCB-Tryck, Örebro 1997.03