Guidelines on the Design and Implementation of Statistical Metainformation Systems

Bo Sundgren

R&D Report
Statistics Sweden
Research - Methods - Development
1993:4

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 1993:4. Guidelines on the design and implementation of statistical metainformation systems / Bo Sundgren.
Digitaliserad av Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-1993-X101OP9304

Guidelines on the Design and Implementation of Statistical Metainformation Systems

Bo Sundgren

R&D Report
Statistics Sweden
Research - Methods - Development
1993:4

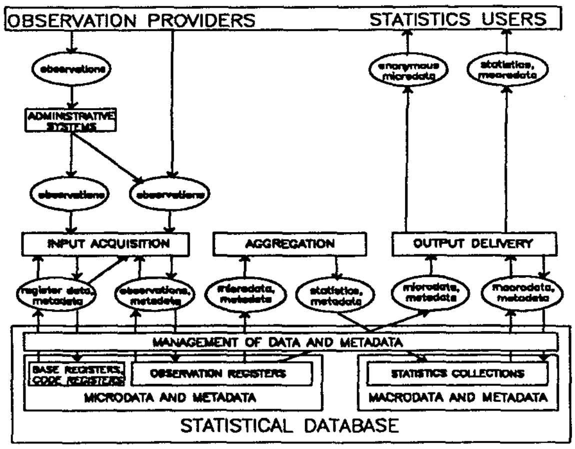
Från trycket	September 1993
Producent	Statistiska centralbyrån, utvecklingsavdelningen
Ansvarig utgivare	Lars Lyberg
Förfrågningar	Bo Sundgren, 08-783 41 48

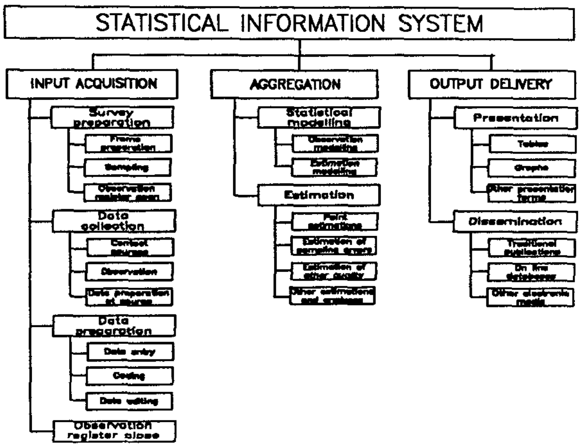
ISSN 0283-8680

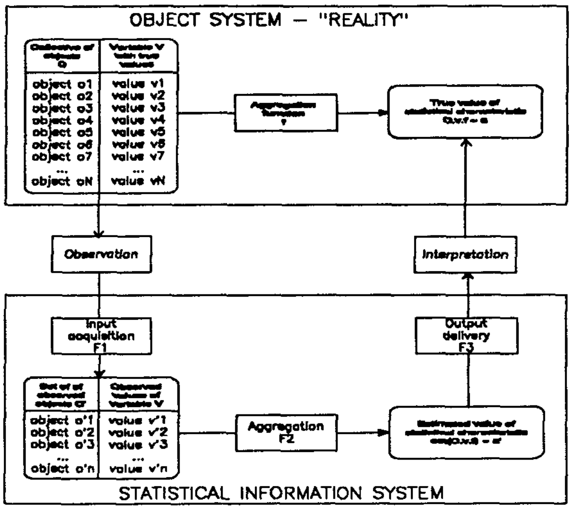
Printed	September 1993
Producer	Statistics Sweden, Department of Research and Development S-115 83 Stockholm
Publisher	Lars Lyberg
Inquiries	Bo Sundgren, telephone 46 08 783 4180

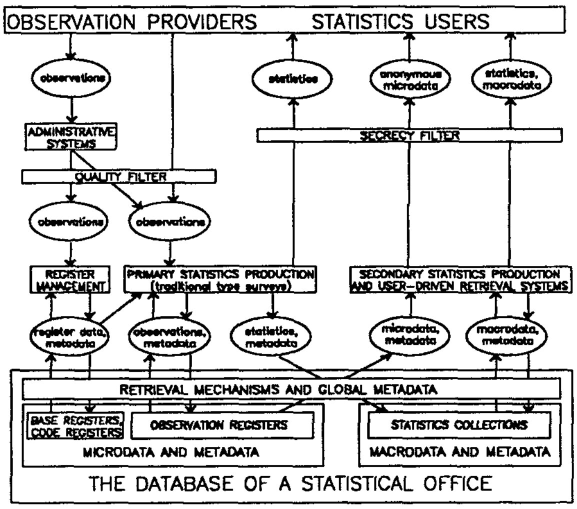
ISSN 0283-8680

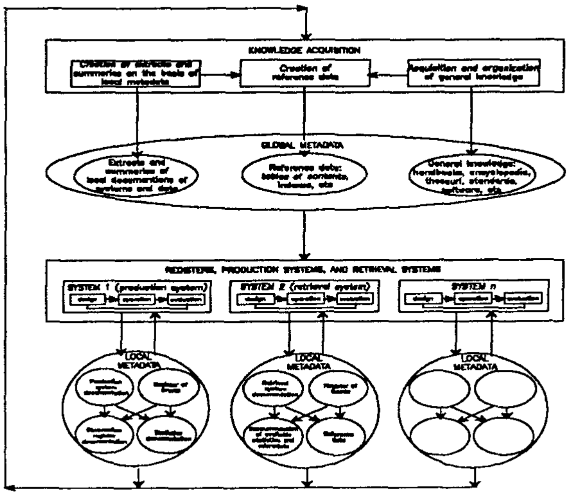
GUIDELINES ON THE DESIGN AND IMPLEMENTATION OF STATISTICAL METAINFORMATION SYSTEMS

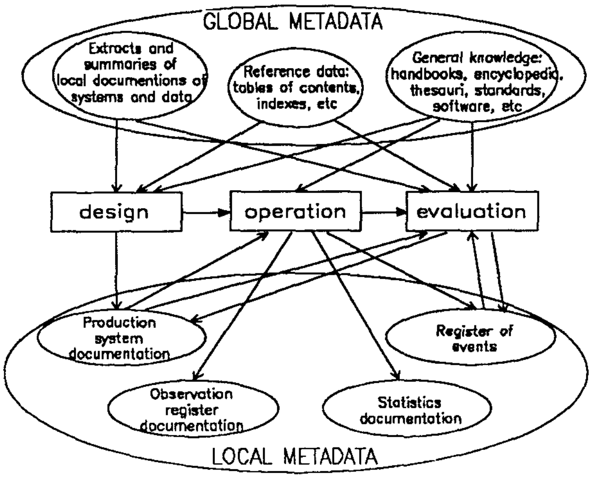









KNOWLEDGE ACQUISITION
Creation of extracts and summaries on the basis of local metadata
Creation of reference data
Acquisition and organization of general knowledge
GLOBAL METADATA
Extracts and summaries of local documentations of systems and data
Reference data: tables of contents, indexes, etc
General knowledge: handbooks, encyclopedia, thesauri, standards, software, etc
REGISTERS, PRODUCTION SYSTEMS, AND RETRIEVAL SYSTEMS
SYSTEM 1 (production system)
SYSTEM 2 (retrieval system)
SYSTEM n
design
operation
evaluation
LOCAL METADATA


GLOBAL METADATA
Extracts and summaries of local documentions of systems and data
Reference data: tables of contents, indexes, etc
General knowledge: handbooks, encyclopedia, thesauri, standards, software, etc
design
operation
evaluation
Production system documentation
Register of events
Observation register documentation
Statistics documentation
LOCAL METADATA

Bo Sundgren

August 1993

Table of contents

Purpose and structure of the *Guidelines*

Many statistical offices and international organizations are planning and developing statistical metainformation systems. Several international meetings have been devoted to this topic and have attracted great interest. It seems appropriate at this stage to collect some methodological ideas and practical experiences and put them into a uniform framework, which can serve as a handbook and source of inspiration for those who are engaged in the design and implementation of statistical metainformation systems all over the world. These *Guidelines* presented here are intended as a first attempt to establish a framework of the kind mentioned.

The material is organized in four chapters. *Chapter 0* introduces a number of basic concepts, which are of fundamental importance for a precise understanding of statistical metainformation systems. Many of the concepts needed are concepts, which are well-known (and sometimes even defined) in the general theory and methodology of statistical surveys and information systems.

Chapter 1 analyzes, who the users of statistical metainformation systems are, and for what purposes these users need statistical metainformation and related services. The analysis is general, but a pattern of analysis is presented (in the form of a specification templet), which could be used in specific practical situations as well.

Chapter 2 goes deeper into the needs for metainformation and metainformation-related services that were identified for different user categories in chapter 1. Several specification and documentation templets are introduced, and once again these templets are believed to be useful in specific practial situations as well as for the general analysis presented in these *Guidelines*. A conceptual model is developed, which explains and gives an overview of the different kinds of metainformation holdings and related functions, which are needed in a statistical metainformation system.

If the first chapters are design-oriented, the final chapter of the *Guidelines*, *chapter 3*, focuses on the implementation problems. It discusses, *inter alia,* how to strike a reasonable balance between important short-term priorities (like orientation towards the needs of statistics users) and the long-term requirements for completeness, consistency, and rationality. It is suggested that a modern statistical office should implement a metainformation infrastructure, which lives in close symbiosis with the kernel operations of the office. Ideally, the metadata should be as automatically captured and transformed as the data which they describe. There are interdependencies between the metainformation flows in several dimensions, for example in the life-cycle dimension (design, operation, evaluation) and between different kinds of statistical information systems (registers, production systems, retrieval systems, etc). For the sake of quality and economy of operation, it is essential that these interdependences are carefully considered, when a statistical metainformation infrastructure is implemented. Finally, some recommendations in the form of a "master plan" for the development of a statistical metainformation infrastructure are given.

The author of this document would like to stress its preliminary status. Constructive criticism of the proposed framework, as well as ideas and experiences to be added to a revised version of the *Guidelines*, are most welcome.

0 Basic concepts

0.1 *Statistical metainformation systems*

According to a simple definition a **statistical metainformation system** is a system, which provides information and information-related services concerning a statistical information system or a system of such systems.

This definition explains *what* statistical metainformation systems are, but not *why* they are needed. A more purpose-oriented definition is that a statistical metainformation system is an information system, which helps people to design, operate, use, and evaluate statistical information systems.

It is obvious from both definitions that in order to understand, what a statistical metainformation system is, we need a rather precise understanding of what a statistical information system is.

0.2 Statistical information systems

A **statistical information system** can be defined as a system which provides statistical information and related services concerning a certain "reality" - the **object system** - to **statistics users**. The statistics users are assumed to have tasks, which imply needs to

- describe and understand the object system and its subsystems and components; and/or

- plan, implement, monitor, and evaluate decisions and actions vis-à-vis the object system.

A statistical information system accomplishes its tasks by performing three major functions:

(F1) an **input acquisition function**, which directly and/or indirectly **observes** (measures) certain **object system characteristics**, and which **prepares and stores** the information thus obtained in the form of data, so-called **microdata**;

(F2) an **aggregation function**, which transforms the microdata produced by the input acquisition function into **macrodata**, or **"statistics"**, which are **estimated values of statistical characteristics** (see definition of *"statistical characteristic"* in section 0.3 below);

(F3) an **output delivery function**, which makes macrodata (statistics) available to the users, and which assists the users to interpret and analyze the data further.

Figure 0.1 illustrates a model of a statistical information system, which contains the three major functions. Since the statistical information system is assumed to be **database-oriented** and **self-describing** (cf below), the model also contains an auxiliary function for the management of data and metadata.

Figure 0.2 illustrates the result of a deeper functional analysis of a statistical information system. This model will be used as a basis for documentation templets proposed in chapter 2.

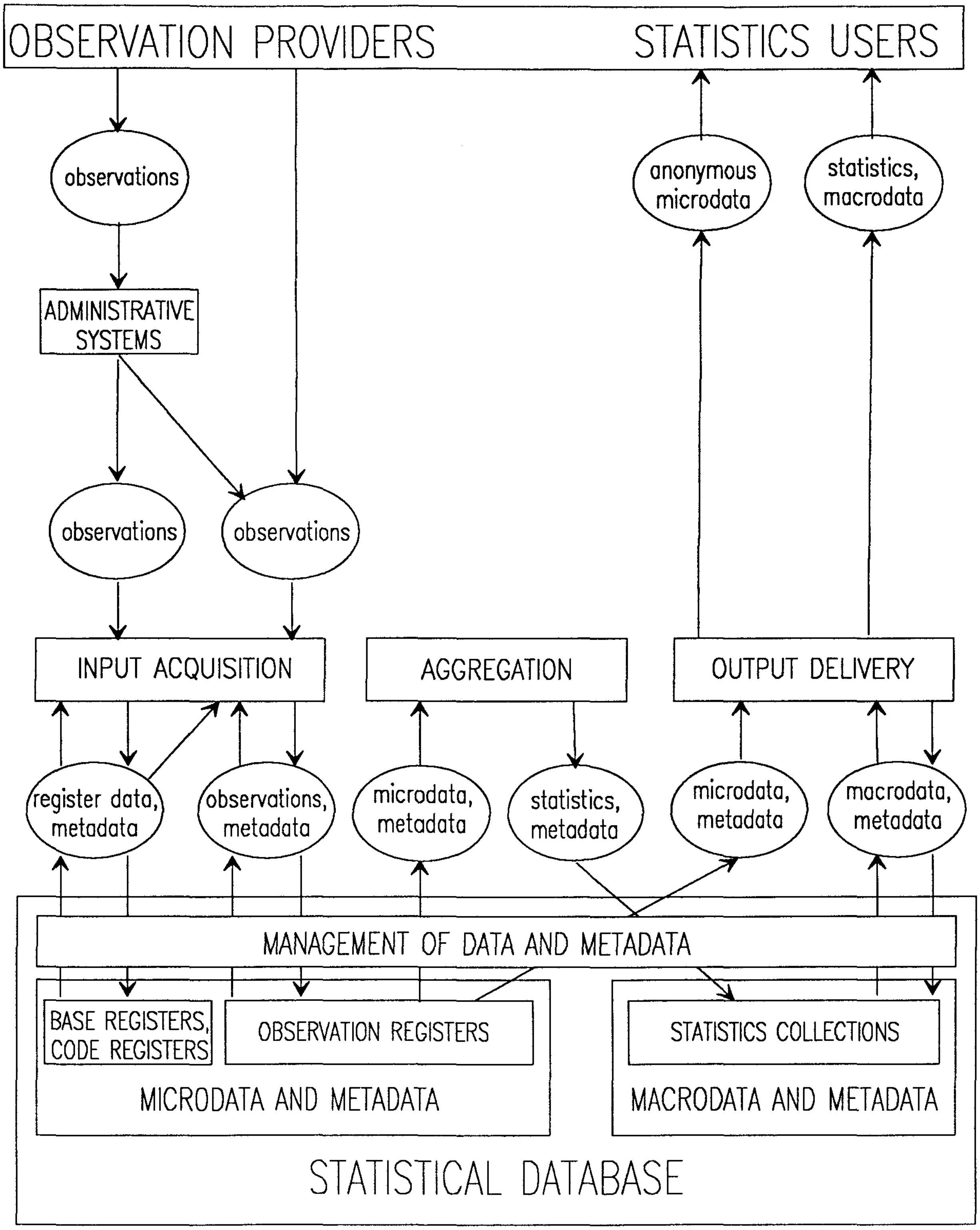


Figure 0.1. *A model of a self-describing database-oriented statistical information system. (The model will be further discussed and elaborated in chapter 3; cf figure 3.1.)*

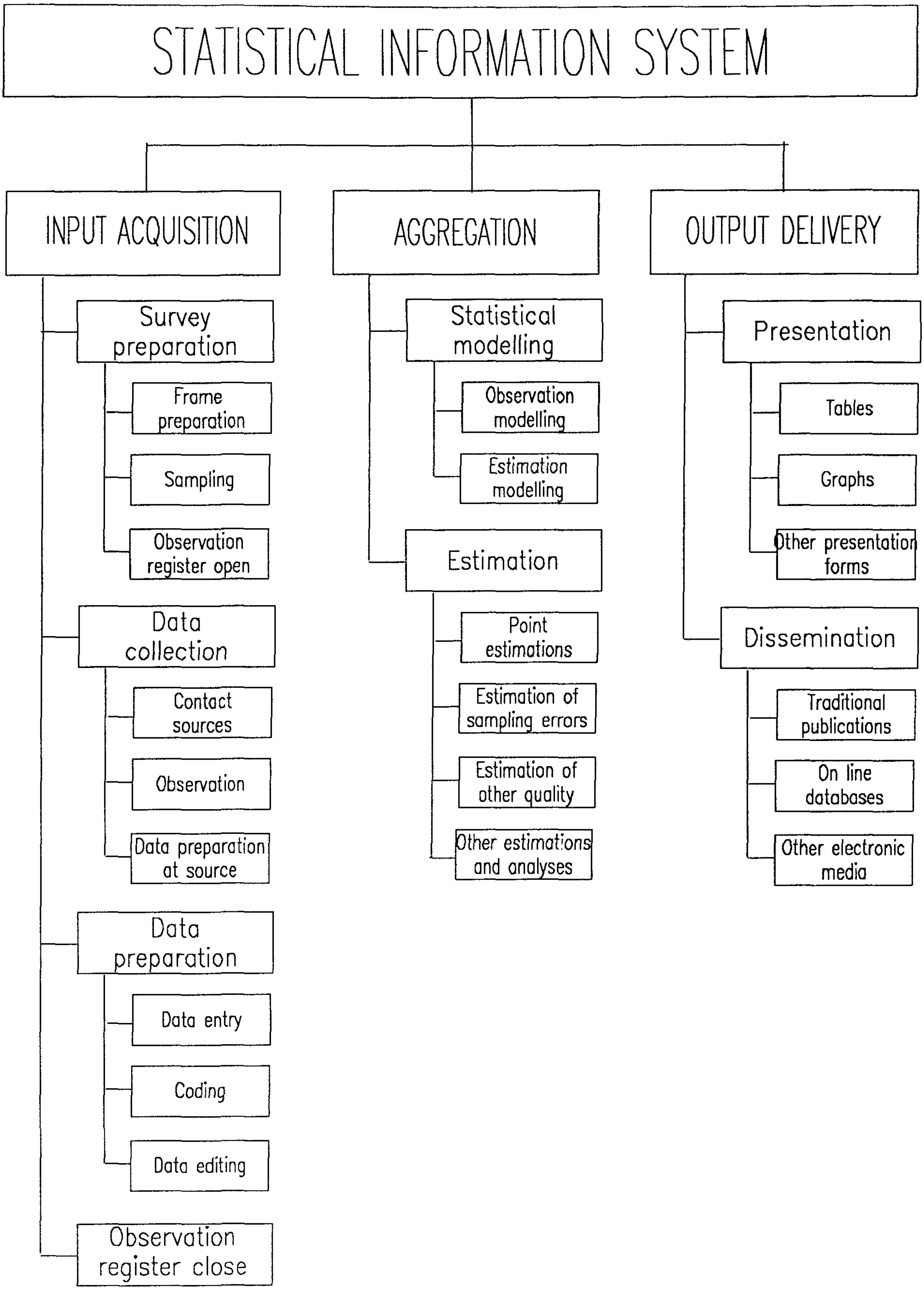


Figure 0.2. *A functionally oriented model of a statistical information system.*

A simple example of a statistical information system is a "traditional" **statistical survey**. Another type of statistical information system, which statistical offices are now paying much attention to, are **retrieval systems**. A statistical survey is focusing on a certain data collection process, resulting in a certain collection of microdata, which are aggregated into estimated values of certain statistical characteristics. In contrast, a retrieval system focuses on the needs of a particular category of statistics users, and aims at making available macrodata and microdata from different surveys (and other sources), which may be relevant for the particular category of users.

In addition to statistical surveys and retrieval systems, there are other types of statistical information systems of a more auxiliary nature. One example is **registers** (cf figure 0.1). There are two kinds of registers, which are particularly important for statistical information systems: **base registers** and **code registers**. A **base register** establishes and maintains an authorized list of the objects belonging to a certain population. A **code register** establishes and maintains an authorized list of the values belonging to the value set of a certain variable or classification.

A complex statistical information systems may contain many statistical surveys, retrieval systems, registers, etc, as subsystems. "The statistical information system of a country" is an example of such a complex statistical information system.

In a **database-oriented statistical information system**, the microdata and macrodata, which are stored and processed by the three major functions (input acquisition, aggregation, output delivery), are communicated within and between the functions via a database.

In a **self-describing statistical information system**, the microdata and macrodata are described by means of accompanying **metadata**, which are stored in the database, and which are consistently transformed, whenever the described data are transformed.

0.3 Statistical characteristics

A **statistical characteristic** is a **property** of a **collective of objects** in the object system. Furthermore, this property, or **variable**, of the collective of objects is a well-defined **function** of one or more properties (variables) of the individual objects in the object collective. Thus a statistical characteristic can be formally described as a triple

(0.1) $C = <O, V, f> = O.V.f$

where

(i) O is an **object collective**, that is, a set of objects - often called a **population**;

(ii) V is a **vector of variables** (often one single variable), each of which have values for the objects in O, usually one value per object and time value;

(iii) f is a function, a so-called **aggregation function** (like frequency *count*, *sum*, *average*, *mean*, *correlation*, *variance*, etc); the aggregation function is defined to provide the **true value of the statistical characteristic**, when it operates on the true values of the variables in V for the objects in O.

Note 1. O.V.f is a useful **dot notation** for the triple <O, V, f>. The dot notation is used in a language called INFOL for describing and manipulating statistical information.

Note 2. If a variable has exactly one value per time value for all objects in O, then it is a **single-valued** variable. If a variable has more than one value per time value for some objects in O, then it is a **multi-valued** variable. If a variable has no value at all for some objects in O, then the variable is only **partially relevant** for the object collective O.

0.4 Statistical information and statistical data

The term **"statistics"** usually denotes macrodata only, that is "estimated values of statistical characteristics", whereas **"statistical data"** often denotes *both* macrodata *and* the microdata, which are used as input to the aggregation process producing the macrodata.

Statistical data are representations of **statistical messages,** which inform about estimated statistical characteristics and underlying observations of object characteristics. Macrodata are representations of **statistics messages,** or **s-messages.** Microdata are representations of **observation messages, o-messages.**

An **s-message** must somehow provide

(i) a *reference to* an **object collective,** $O(t_0)$, which is well-defined in time and space;

(ii) a *reference to* a **vector of variables,** $V = <V(t_1), ..., V(t_n)>$, which have well-defined but usually unknown values for the objects in the object collective at certain specified points or intervals of time, $t_1, ..., t_n$, respectively;

(iii) a *reference to* an **aggregation function,** f, which is well-defined for V;

(iv) a **value,** which is the **estimated value of the statistical characteristic**

(0.2) $C = <O(t_0), V(t_1, ..., t_n), f> = O(t_0).V(t_1, ..., t_n).f$

Note. An **estimated value** c' of a statistical charachteristic C will typically be different from the **true value** c. The discrepancy is due to errors and uncertainties of different kinds; see section 2.2.

An **o-message** must somehow provide

(i) a *reference to* an **object collective,** $O(t_0)$, which is well-defined in time and space, and to which the **observed object** belongs; the observed object may be **identified** or **anonymous;**

(ii) a *reference to* a **vector of variables,** $V = <V(t_1), ..., V(t_n)>$, the **observed variables,** which are supposed to have well-defined true values for the objects in the object collective at the specified points or intervals of time, $t_1, ..., t_n$, respectively;

(iii) a vector of values, $v' = <v'_1, ..., v'_n>$,which are the observed values of V; naturally the observed values may be different from the corresponding true values $v = <v_1, ..., v_n>$.

Figure 0.3 illustrates the fundamental concepts introduced in sections 0.3 and 0.4..

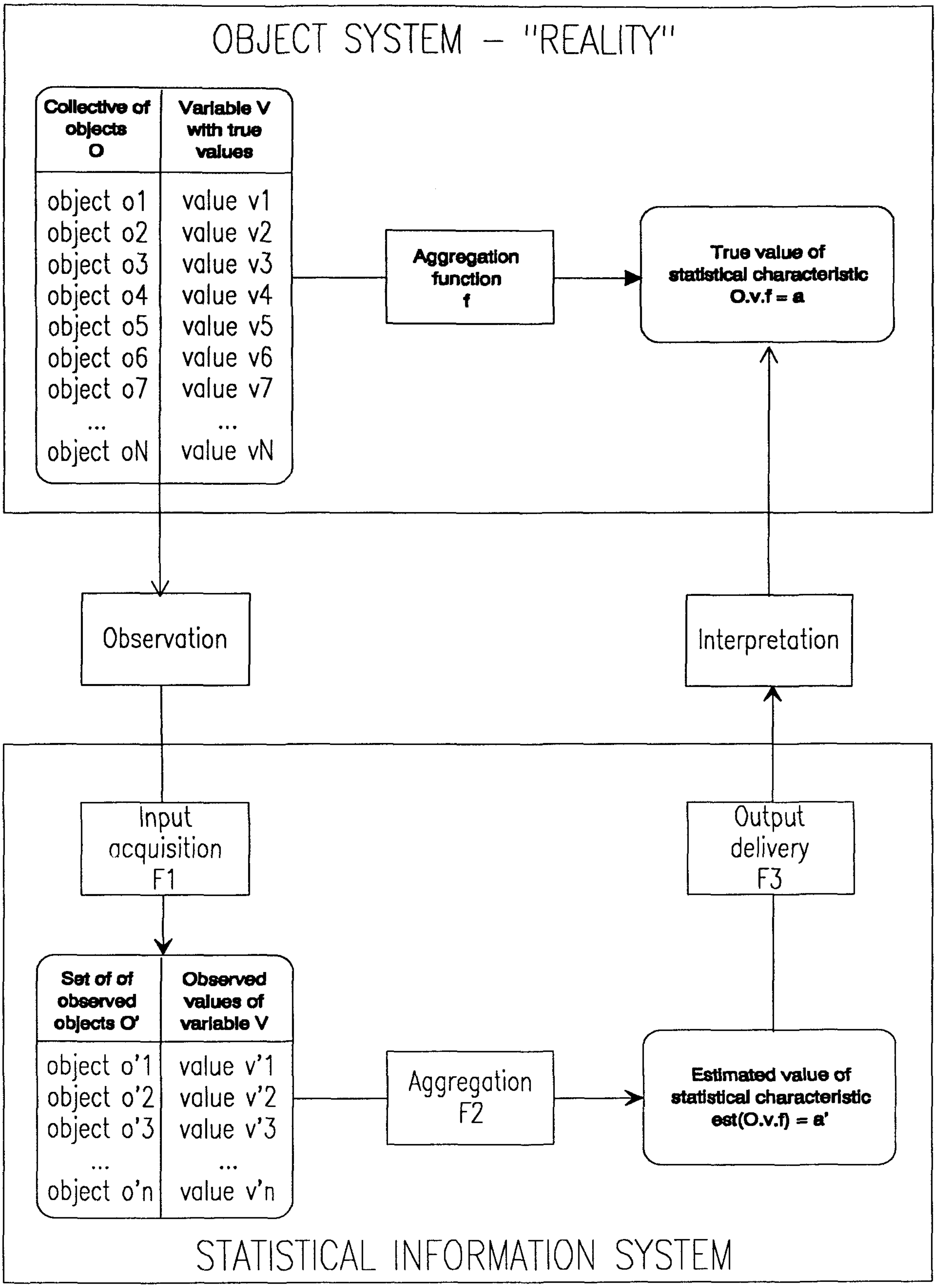


Figure 0.3. *Illustration of some fundamental concepts in statistical information processing.*

0.5 Structured sets of statistical characteristics and statistical data

Statistical data, particularly macrodata, are often organized in certain typical **structures.** Thus, for example, statistics users are often interested to obtain estimated values of "the same" statistical characteristic for

- a series of time periods (rather than a single one) - **"time series data"**; and/or

- a structured set of object collectives (rather than a single one) - **"structured cross-section data".**

Time series data may be indicated as such by using a **time parameter** as part of the names of the object collectives and variables, which are part of the statistical characteristic. For example, data labeled something like "average income for persons 1991, 1992, and 1993" may be formally described in the following way:

(0.3) PERSON(t_p).income(t_i).average;

where

t_p = 1991-01-01, 1992-01-01, 1993-01-01;

t_i = the year starting at $t_{p.}$

The collective of objects, or **population**, referred to in the definition of a statistical characteristic, is often subdivided into sub-collectives, sometimes called **domains of interest**. Such a structured set of related object collectives may be called a **structured population.** The structuring is often accomplished by means of **crossclassification**, using the Cartesian product of the value sets of a number of variables. For example, a structured population labeled

(0.4a) "persons by occupation, age_group, and sex 1991"

may be formally described as:

(0.4b) PERSON(1991)(by occupation(1991) * age_group(1991) * sex(1991));

If we combine the structuring mechanisms of (0.2) and (0.3) we get a rather complex time series of cross-classified cross-sectional data. For example, a set of statistical data labeled

(0.5a) "average income for persons by occupation, age_group, and sex 1991, 1992, and 1993"

may be formally described as

(0.5b) PERSON(t_p)(by occupation(t_p) * age_group(t_p) * sex(t_p)).income(t_i)average;

where

(i) t_p = 1991-01-01, 1992-01-01, 1993-01-01;

(ii) t_i = the year starting at $t_{p.}$

A generalization of this format for specifying structured sets of statistical characteristics and statistical data is the following **box structure** format, which is borrowed from the earlier mentioned language INFOL:

(0.6) <object type>(t_0)[(with <property>)]
[(by $vg_1(tg_1)$ * ... * $vg_n(tg_n)$).
(($vb_1(tb_1)$, ..., $vb_m(tb_m)$).<aggregation function>;

where

(i) <object type> denotes a **time-independent** object collective, which is made **time-dependent** by means of the qualifier (t_0), which may be either a parameter (in the case of a time series) or a constant;

(ii) the optional clause [(with <property>)] indicates a **selection** of a subset of <object type>(t_0) by means of <property>, which may be expressed in terms of variables variables, which are defined and relevant for the objects in the object collective;

(iii) the two clauses described in (i) and (ii) form a part of the box structure, which is sometimes referred to as the **alfa part**;

(iv) the value sets of the so-called **gamma variables**, $vg_1(tg_1)$, ..., $vg_n(tg_n)$, crossclassify the time-dependent object collective;

(v) the gamma variables are **time-qualified**, and any one of the qualifiers may be either a parameter (time series case) or a constant;

(vi) the variables $vb_1(tb_1)$, ..., $vb_m(tb_m)$ are the so-called **beta variables**, which are also time-qualified (by means of parameters and/or constants), and the values of which are aggregated by means of <aggregation function>.

Note 1. In practice, many of the time parameters (time constants) occurring in a box structure expression are actually the same, and then they may be separately specified in a special time clause, sometimes referred to as the **tau part**.

Note 2. Box structures following the format given above are sometimes referred to as **alfa-beta-gamma-tau-structures**.

1 Users and usages of statistical metainformation systems

The purpose of a statistical metainformation system is to provide information and information-related services concerning a statistical information system. One way of breaking down this rather general description of the purpose of a statistical metainformation system into a number of more specific task descriptions is to identify and analyze the different activities, or processes, which make up the life cycle of a statistical information system. For each idenfied activity we could try to identify the most important actors and the most important tasks of those actors, which may require metainformation and metainformation-related services. This is what we are going to do in this section of the *Guidelines*.

1.1 Major phases of the life cycle of a statistical information system

The life cycle of a statistical information system consists of the following major phases (cf figure 1.1):

(P1) a process leading to a decision to develop a statistical information system;

(P2) a development phase, consisting of design and implementation activities;

(P3) an operation and maintenance phase, involving users and producers;

(P4) an evaluation process, possibly leading to a decision to redesign or discontinue the statistical information system;

(P5) a dismantling phase.

The five major phases of the life cycle of a statistical information system are to some extent overlapping. For example, some preliminary design activities (and even implementation activities) will often be a natural part of the initial decision process leading to the decision to (or not to) develop a statistical information system, and some evaluation activities will most likely go on more or less continuously during the operation phase.

Phase P1 is a **management process**, involving decision-makers, "managers" on a relatively high level. Considerable resources, financial and others, are usually required to develop a statistical information system, and such resources can only be committed by high-level decision organs. Other actors involved in phase P1 are designers and (representatives of) the future users and producers involved in the operation of statistical information system.

Phase P2, the **development phase**, is dominated by people, who are responsible for the design and implementation of the statistical information system. These designers can be subdivided into three categories: subject matter specialists, statistical methodologists, and information system specialists. It is important to note, however, that the development phase has to be appropriately controlled by authorized representatives of users and producers and led by professional managers.

Phase P3, the **operation and maintenance** phase, can be seen from two alternative and complementary perspectives: the **user perspective**, and the **producer perspective**. Traditionally, the producer perspective has dominated in statistical offices, not so much in the sense that the needs of statistics users have been disregarded, but rather in a more structural sense: statistical information systems have been formed around a single survey or a small number of surveys, which are mainly related by input and production oriented constraints.

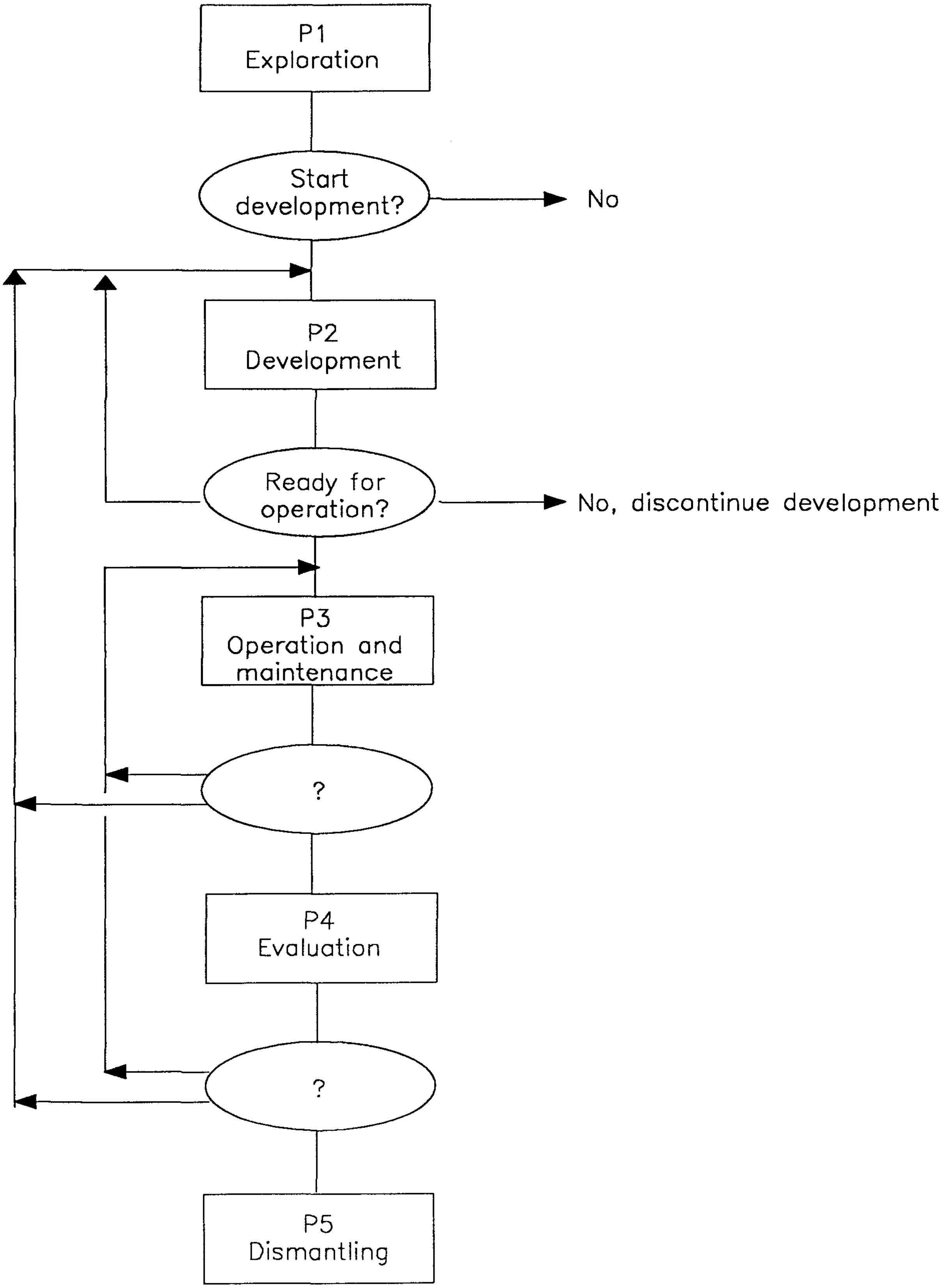


Figure 1.1. *Five major phases in the life cycle of a statistical information system.*

Modern statistical offices as well as international statistical organs and, of course, the more and more competent and active statistics users are now rightfully stressing the need to balance the traditional producer-oriented perspective with a perspective dominated by user needs.

The term "**maintenance**" is used for (minor) adjustment activities, which fall in the borderland between "operation" and "redesign". Once again, maintenance activities can be motivated (or even necessiated) by either user/output-oriented or production/input-oriented factors.

Phase P4, **evaluation**, is another management process. It is often spontaneously activated by the users and producers themselves, but from time to time it is important for the managers responsible for the statistical information system under consideration, as well as for higher-level managers with a wider responsibility, to initiate more systematical evaluations. There are three possible outcomes of such an evaluation:

- a decision to redesign the statistical information system to a greater or lesser extent;
- a decision to discontinue the operation of the statistical information system;
- a decision to continue the operation as before, without any redesign.

Phase P5, the **dismantling** of a statistical information system, is often neglected in the sense that it is not explicitly managed and operated. This may lead to an unnecessary waste of resources and to losses of valuable experiences.

1.2 Statistical information system actors - potential users of a statistical metainformation system

To summarize the analysis in section 1.1, we have identified the following categories of actors, who are involved in the major phases of the life cycle of a statistical information system:

(U1) **users** of the information outputs and services from the statistical information system;

(U2) **producers** of the information outputs and services;

(U3) **designers**, or rather designers/redesigners/maintainers, of the statistical information system consisting of three subcategories:

- **subject matter specialists**
- **statistical methodologists**
- **information system specialists**;

(U4) **managers**, consisting of the subcategories

- **local managers**, responsible for a limited statistical information system;
- **global managers**, who have a wider responsibility.

These categories of actors in the activities of a statistical information system are the potential users of the information outputs and services of a statistical metainformation system.

Theoretically, a statistical metainformation system could work without computers. Such a system would be manual and more or less informal. However, for all practical purposes we may assume that computers and computerized processes and metadata holdings will be essential components of any modern metainformation system. In fact, software artifacts will usually have such an important intermediary role between the above-mentioned human actors and the subsystems and components of a statistical metainformation system that it is well justified to regard these software artifacts themselves, in their own right, as "users" of metainformation and metainformation-related services. Thus we have a fifth actor/user category to add to the list above:

(U5) metainformation system software components, or **software artifacts.**

Since software artifacts are by definition computerized, the metainformation and metainformation-related routines that they make use of have to be computerized, too, usually in the form of more or less strictly formalized metadata and metadata handling algorithms. Furthermore, it should be noted that the software artifacts themselves, by their way of functioning, generate certain needs for metadata and metadata handling capabilities, which can be precisely defined only during a particular software design or acquisition process.

Figure 1.2 illustrates which actor categories are likely to be the most active and responsible users of (more or less formalized) metainformation and metainformation handling capabilities during the respective phases of the life cycle of a statistical information system.

1.3 Tasks requiring statistical metainformation and related services

The lists of life cycle phases and actors involved may be combined like in figure 1.2 and then used as a starting-point for identifying tasks in statistical information systems where statistical metainformation and metainformation-related services are needed. The specification templet in figure 1.3 shows the result of making such an inventory of tasks.

	(P1) Exploration	(P2) Development	(P3) Operation and maintenance	(P4) Evaluation	(P5) Dismantling
(U1) Users	p	p	p	p	
(U2) Producers	p	p	p	p	p
(U3) Designers -subject matter -statistical method -information system	p	R		p	
(U4) Managers	R	p	R	R	R
-local			R		R
-global	R				
(U5) Software		p	p	p	p

Figure 1.2. *Crosstabulation of actor categories by life cycle phases indicating. "R" indicates responsible actor category, "p" participating.*

1 STATISTICS USERS

1.1 Search for, identify, and locate possibly relevant statistical data, metadata, and analyses

1.2 Make a cost/benefit-analysis and decide what should actually be retrieved

1.3 Specify retrieval requests and carry out retrieval operations

1.4 Interpret and process retrieval outputs

2 STATISTICS PRODUCERS

2.1 Recall and carry out the instructions for operating a certain statistical survey or information system

2.2 Collect operation experiences and propose improvements

2.3 Train new staff members

3 DESIGNERS

3.1 Design a new statistical survey or information system, covering all aspects: subject matter, statistical methodology, and information technology

(a) Specify survey contents and outputs
(b) Develop a survey plan
(c) Design and implement input-oriented routines and final observation register
(d) Design and implement statistical processing and output-oriented routines

3.2 Evaluate and improve an existing statistical survey or information system

4 MANAGERS

4.1 Monitor and evaluate the performance of the managed system (of surveys and systems)

(a) Costs and production experiences
(b) Revenues and user satisfaction

4.2 Encourage proposals for improvements, evaluate proposals, and carry out projects

5 SOFTWARE

5.1 Support the design of a new survey or information system

5.2 Support the operation and maintenance of a statistical survey or information system

5.3 Support the monitoring, evaluation, and redesign of a statistical survey or information system

Figure 1.3. *Specification templet for users and usages of statistical metainformation systems.*

2 Information needs and functional requirements to be met by a statistical metainformation system

The specification templet in figure 1.3 for users and usages of statistical metainformation systems is a good starting-point for identiying the needs for metainformation and metainformation-related system functions to be met by a statistical metainformation system. As designers of such a system we may work ourselves down the tasks listed in the specification templet, and for each task we may ask ourselves which information sets and system functions would be needed for carrying out the task in a satisfactory way. Most information sets and system functions thus identified are likely to be computerized or at least computer-supported, when later implemented (cf section 3 of the *Guidelines*), but in principle we should not at this stage bother about implementation aspects. The specification should be a conceptual one, formulated in terms of abstract information sets and functionalities.

Figure 2.1 shows a possible result of such a specification process. It contains a specification templet for the information contents and functionality required from a statistical metainformation system by its users. It is a specification templet which is meant to be rather generally applicable. When a specific application is at hand, for example, in a specific national statistical office or international statistical organization, the specification templet will have to be extended and modified. However, the general templet shown in figure 2.1 should be a good starting-point for such a more specific and practically oriented exercise.

The bold-typed tasks in the specification templet 2.1 are the same as the tasks listed in the specification templet 1.3. For each task a number of information and/or functionality requirements have been listed with italicized typing.

2.1 Categorization of metainformation holdings and functions

Analyzing the contents of templet 2.1, we may identify some major types of metainformation holdings and functions that a statistical metainformation system would have to have. As illustrated by the specification templet in figure 2.2, we may group the metainformation requirements into two major categories called **specific knowledge** and **general knowledge**, respectively.

Specific knowledge denotes metainformation and metainformation-related functions associated with individual systems, production systems and retrieval systems, that is, metainformation of a relatively local nature, if regarded, for example, in relation to the statistical information system of a certain national statistical office, or even the statistical information system of a world community of some kind.

General knowledge denotes metainformation and metainformation-related functions of a more global character, for example knowledge about how statistical surveys and information systems are to be designed, operated, and evaluated, which we may call **handbook knowledge**, since it is often documented in the form of handbooks, manuals, and guidelines. Another type of general knowledge are **encyclopedical knowledge**, that is, knowledge of the type documented in dictionaries, encyclopedias, and thesauri. Yet another type of general knowledge concerns **standards**, contents-oriented as well as dealing with representation formats. Finally there is a subcategory of general knowledge that is concerned with and contained in **software** products.

1 STATISTICS USERS

1.1 Search for, identify, and locate possibly relevant statistical data, metadata, and analyses

1.1.1 Descriptions - more or less formalized - of available statistics and observation registers

1.1.2 Well-structured and informative tables of contents - as global as possible - specifying available statistics and observation registers and giving references to the more detailed descriptions

1.1.3 Indexes to the tables of contents

1.1.4 Algorithms for automatical indexing and free-text searches

1.1.5 Thesauri for supporting the process of specifying search questions by providing broader terms, narrower terms, and related terms

1.2 Make a cost/benefit-analysis and decide what should actually be retrieved

1.2.1 Precise definitions of the meaning and quality of available micro-level and macro-level message types, including relevant standard definitions

1.2.2 Price lists and lead time information for available statistics and microdata

1.2.3 References to related previous analyses, carried out by other users

1.3 Specify retrieval requests and carry out retrieval operations

1.3.1 Documentation and help information concerning retrieval procedures

1.4 Interpret and process retrieval outputs

1.4.1 Precise definitions of the meaning and quality of available micro-level and macro-level message types, including relevant standard definitions (cf 1.2.1)

1.4.2 References to related previous analyses, carried out by other users (cf 1.2.3)

1.4.3 Encyclopedia of methods for statistical analysis

2 STATISTICS PRODUCERS

2.1 Recall and carry out the instructions for operating a certain statistical survey or information system

2.1.1 Detailed documentation of all processes of the production system, including explanations of the rationale behind the processes and detailed descriptions of inputs and outputs

2.2 Collect operation experiences and propose improvements

2.2.1 Metadatabase (accounting system) for recording and retrieving operation experiences

2.3 Train new staff members

2.3.1 Detailed documentation of all processes of the production system, including explanations of the rationale behind the processes and detailed descriptions of inputs and outputs (cf 2.1.1)

Figure 2.1a. *Specification templet for information contents and functionality required from a statistical metainformation system by its users; continued on next page.*

3 DESIGNERS

3.1 Design a new statistical survey or information system, covering all aspects: subject matter, statistical methodology, and information technology

3.1.1 Handbook in the design of statistical surveys and information systems

3.1.2 Encyclopedia of statistics production methods

3.1.3 Association system for identifying, locating, and retrieving descriptions of "similar" surveys, information systems, and system components

3.1.4 Library of standards: definitions of observation characteristics, statistical characteristics, message types, measurement methods, message representation formats, etc

3.2 Evaluate and improve an existing statistical survey or information system

3.2.1 Handbook, encyclopedia, association system, and library of standards (cf 3.1.1-4)

3.2.2 Documentation database and accounting system covering the system under consideration as well as "similar" systems

4 MANAGERS

4.1 Monitor and evaluate the performance of the managed system (of surveys and systems)

4.1.1 Accounting system covering the managed system (of statistical surveys and information systems) as well as "similar" or otherwize related systems (cf 3.2.2)

4.2 Encourage proposals for improvements, evaluate proposals, and carry out projects

4.2.1 Communication system (e-mail, conference system, etc)

4.2.2 Handbook in project evaluation and project management

5 SOFTWARE

5.1 Support the design of a new survey or information system

5.1.1 Algorithms and data sets required by software supporting and partially automating the design process (cf 3.1)

5.2 Support the operation and maintenance of a statistical survey or information system

5.2.1 Algorithms and data sets required by software supporting and partially automating the operation of a statistical survey or information system, as seen from (a) a user perspective, and (b) a producer perspective (cf 1 and 2, respectively)

5.2.2 Algorithms and data sets required by software supporting and partially automating the maintenance of a statistical survey or information system (cf 3.2)

5.3 Support the monitoring, evaluation, and redesign of a statistical survey or information system

5.3.1 Algorithms and data sets required by software supporting and partially automating the monitoring, evaluation, and redesign of statistical surveys and information systems (cf 3.2 and 4)

Figure 2.1b. *Specification templet for information contents and functionality required from a statistical metainformation system by its users; continued from previous page.*

SPECIFIC KNOWLEDGE

S1 Survey-related knowledge

S1.1 Documentation of production systems and statistical data

S1.1.1 Production system documentation
S1.1.2 Observation register documentation
S1.1.3 Statistics documentation
S1.1.4 Analysis documentation

S1.2 Register of survey-related events

S1.2.1 Log of survey design changes
S1.2.2 Log of production system events
S1.2.3 Accounting system

S2 Retrieval-related knowledge

S2.1 Retrieval system documentation

S2.2 Reference data: tables of contents, indexes etc

S2.3 Register of retrieval-related events, including accounting system

GENERAL KNOWLEDGE

G1 Handbook knowledge

G1.1) Handbook in the design of statistical surveys and information systems

G1.2 Handbook in project evaluation and project management

G2 Encyclopedical knowledge

G2.1 Encyclopedia of statistics production methods

G2.2 Dictionary of terms used by designers of statistical surveys and information systems

G2.3 Thesaurus of subject-matter oriented statistical terminology

G3 Standards

G3.1 Contents-oriented standards

G3.1.1 Definitions of observation characteristics (object types, variables, values, etc), including measurment methods and instruments
G3.1.2 Definitions of statistical characteristics

G3.2 Representation-oriented standards

G3.3.1 Message representation formats
G3.3.2 Tabulation layouts

G4 Software and software-related knowledge

Figure 2.2. *Specification templet for information sets and related functions to be contained in a statistical metainformation system.*

2.2 Documentation of published statistics

A printed publication of some sort is the traditional form of appearance for aggregated statistical data, **statistics**. Today there are several alternative forms for presenting and publishing statistics, for example on-line databases and CDROM disks. A general documentation templet for statistics should be applicable to all kinds of aggregated statistical data, regardless of their form of appearance. However, since most users and producers of statistics are still used at thinking of statistics in the form of traditionally published statistics, we shall start our analysis of metainformation about statistics from that perspective.

Figure 2.3 shows a **documentation templet for published statistics.** It could also be seen as a **quality declaration** templet.

The three main parts (parts 1, 2, and 3) of the documentation templet for published statistics in figure 2.3 aim at answering the three basic questions that a statistics user will ask with regard to a collection of statistics:

1. *What is the meaning of the statistics?* (The **relevance** aspects.)
2. *How accurate are the statistics?* (The **accuracy** aspects.)
3. *How can the statistics be retrieved and used?* (The **availability** apects.)

In addition, the documentation templet contains a part 0 and a part 4. Part 0 contains some administrative information, as well as references to the surveys underlying the published statistics here documented. Part 4 contains a register, or a so-called **log-book**, of important design changes and production/retrieval events that have occurred visavi the documented (time series of) statistics over time.

As already indicated, the answers to the questions 1, 2, and 3 above could be seen as a quality declaration of the statistics described. This view implies a **quality concept** consisting of three main dimensions: relevance, accuracy, and availability. Sometimes a more narrow quality concept is used, focusing on the accuracy aspects only, and in particular on those accuracy aspects (sampling errors etc) which are relatively easy to quantify. However, in these *Guidelines* we shall use the broader quality concept.

Some of the concepts occurring in the documentation templet in figure 2.3 will now be commented upon in more detail.

2.2.1 Universe of interest and target universe

A **universe of interest** is a problem area of some kind, which a user of statistics is interested in, and which he or she wants to gets illustrated by means of statistical information. If the statistics user is a politician or an adviser to a politician, the universe of interest could be a certain sector of the economical and/or social life in a country. If the user is a businessperson, it could be the potential market for a certain product.

When the surveys underlying the statistics documented were originally designed, a certain universe of interest was usually conceptualized. However, since the needs of several different users and usages of the planned statistics would typically have had to be taken into account, the design would often result in a **constructed universe of interest**, which would be some kind of compromise between the needs of different potential statistics users, as well as a compromise between what would ideally be desirable and what had to be settled for when practicalities and costs were taken into consideration. This *"compromise universe of interest"* is called the **target universe**. It is the universe, which the statistics were designed to illustrate.

0 Documentation structure etc

- 0.0 Documentation templet
- 0.1 Statistics described
- 0.2 Underlying surveys and information systems
- 0.3 Responsibility: organization, unit, person

1 Relevance

- 1.1 Universe of interest and target universe, verbal description
- 1.2 Target universe, formal description
  - 1.2.1 Target objects
  - 1.2.2 Target populations and domains of interest accounted for in the statistics
  - 1.2.3 Target variables
  - 1.2.4 Structured summary of estimated statistical characteristics, including indications of "degree of detail"
- 1.3 Time
  - 1.4.1 Frequency of the statistics
  - 1.4.2 Comparability over time
- 1.4 Comparability with other statistics

2 Accuracy

- 2.1 Uncertainty contributions from different error sources
  - 2.1.1 Sampling
  - 2.1.2 Non-response
  - 2.1.3 Measurement/observation
  - 2.1.4 Frame coverage
  - 2.1.5 Processing
  - 2.1.6 Model assumptions
- 2.2 Summary description of the total error

3 Availability

- 3.1 Presentation and distribution forms
  - 3.1.1 Printed outputs
  - 3.1.2 Electronical outputs, including databases
- 3.2 Timeliness
  - 3.2.1 Planned and actual delay between status/event and observation/measurement/registration
  - 3.2.2 Planned and actual delay between status/event and publication of statistics
- 3.3 Metadata
  - 3.3.1 Documentation
    - 3.3.1.1 Integrated with the statistics
    - 3.3.1.2 Separately available (including documentations of underlying surveys and systems
  - 3.3.2 Contact persons
  - 3.3.3 Available analyses of the statistics
- 3.4 Possibilities to access and analyse microdata underlying the statistics
- 3.5 Prices and other retrieval conditions

4 Log-book of changes, production, and usage

- 4.1 Major changes, including changes in underlying surveys
- 4.2 Production events and costs
- 4.3 Retrieval events and revenues

Figure 2.3. *Documentation templet for published statistics.*

Naturally, an actual user of some published statistics will have his/her own particular universe of interest, which may differ from the target universe as well as from the universes of interest, which were considered at design time. Such a user will have to judge the relevance of the statistics documented with regard to his/her problem, by comparing his/her universe of interest with the target universe, which the statistics were designed to inform about.

A statistical universe of the kind discussed here (a universe of interest or a target universe) is usually more formally modelled in terms of (macro-level) **statistical characteristics**, which in turn are modelled as derivable from (micro-level) **object characteristics**. As stated in section 0.3, a **statistical characteristic** can be formally described as a triple

(2.1) $C = <O, V, f>$ or, with dot notation, $C = O.V.f$

where

(i) O is an **object collective**;
(ii) V is a **vector of variables** which have values for the objects in O;
(iii) f is a so-called **aggregation function.**

Macrodata consists of **estimated values of statistical characteristics.** The estimation process leading to an estimated value of a statistical characteristic starts from a set of observations of one or more (micro-level) **object characteristics.**

An **object characteristic** can be formally described as a pair

(2.2) $C = <O, V> = O.V$

where

(i) O is an **object collective**; and
(ii) V is a **vector of variables**, which are defined for the objects in O.

The observed set(s) of objects, $O_{o,i}$, and the observed vector(s) of variables, $V_{o,i}$, involved in the observed (micro-level) object characteristics needed for the estimation process do not necessarily have to be the same as the target set of objects, $O_{t,k}$, and the target vector of variables, $V_{t,j}$, involved in the (macro-level) target characteristic, the value of which is to be estimated. However, in practice, there is a rather common special case, where there is only

(i) one set of observed objects, O_o;
(ii) one observed variable, V_o;
(iii) one set of target objects, O_t; and
(iv) one target variable, V_t;

and where

(v) O_o is (by survey design) aimed at being
- either (in the case of a complete enumeration survey) identical with O_t;
- or (in the case of a sample survey) a sample of O_t;

and

(vi) V_o is (by survey design) aimed at being identical with V_t

In this simple special case the estimation process may be formally described as a function e producing estimated values of the target characteristic from observed values of the observation characteristics:

(2.3) $\text{est}(C_t) = e(\text{obs}(C_o))$

where

(i) $C_t = <O_t, V_t, f>$ is the target (macro-level) characteristic;
(ii) est(x) is the estimated value of x;
(iii) $C_o = <O_o, V_o>$ is the observed (macro-level) characteristic;
(iv) obs(y) is a set of observed values of y;
(v) e is a function, which according to statistical theory can be used for estimating f.

Example 2.1. Estimating the value of a population average of a variable on the basis of observed values of variable for a random sample of objects belonging to the population.

A more general situation is characterized by

(i) one set of target objects, O_t;
(ii) one target variable, V_t;
(iii) one or more observed set(s) of objects, $O_{o,i}$, possibly different from O_t;
(iv) one or more observed vector(s) of variables, $V_{o,i}$, possibly different from $V_{t;}$

In the more general type of situation the target characteristic is somehow derived from the observation characteristics, sometimes via intermediate characteristics, where either the object component, or the variable component, or both, are different from the corresponding components of either the observed characteristics, or the target characteristics, or both.

Example 2.2. Estimating the size of a population of objects of a certain type on the basis of observations of birth, death, and migration events, where objects belonging to the population have been involved.

Example 2.3. Estimating the total disposable income of a population of households on the basis of observations of salaries received, taxes paid, etc, by a number of individual persons belonging to households in the target population.

As indicated by item 1.2.4 in figure 2.3, an overview of the target universe should be given by means of a "structured summary of estimated statistical characteristics". A practical and precise formalism for giving this "structured summary" is to use the **alfa-beta-gamma-tau-structuring** described in section 0.5.

2.2.2 Comparability in time and space: direct and indirect descriptions

A statistics user is often interested to know, how comparable a particular estimated value of a particular statistical characteristic - that is, a particular "statistical figure" - is with other "figures" in the same time series, or with similar "figures" for other countries, branches of industry, etc. We refer to this quality aspect as **comparability in time and space**.

There are in principle two different ways of describing comparability, one direct and one indirect way. In order to describe comparability directly, one must foresee, which concrete comparisons a potential user will be interested to make. This may be relatively easy to make for comparability in time, but more difficult for comparability in space, at least if one has the ambition to be reasonably "complete".

An indirect way of describing comparability in time and space is to relate the documented characteristics with standards, for example by stating that a certain characteristic has been defined and measured in accordance with a certain international standard, or by describing how the definition, measurement, etc, deviates from such a standard.

2.2.3 Errors and accuracy

Accuracy indicates how well the actually obtained estimated values of certain statistics coincide with the true values of the statistics, that is, the values that one would have obtained if there had not been any **errors or uncertainties** at all.

It is often difficult to describe accuracy directly. It is usually easier to give an indirect description by describing (if possible quantitatively) the errors and uncertainties caused by different sources. The most important **sources of errors and uncertainties** are listed in item 2.1 of the documentation templet for published statistics in figure 2.3: sampling, non-response, measurement/observation, frame coverage, processing, model assumptions. The documentation templet also encourages an attempt (item 2.2) to make a summary description of the **total error**, for example by indicating the relative importance of the different sources and uncertainties.

2.2.4 The existence of metadata as a quality component

We obviously describe the quality of statistics by means of metadata. However, it is interesting to note a double role of metadata: it is not only used for describing quality; the existence of metadata (that is, quality descriptions) is in itself a quality component. This conclusion can be drawn from item 3.3 in the documentation templet of published statistics in figure 2.3.

2.3 Documentation of stored macrodata sets

In addition to traditional printed publications a modern statistical office will use other forms and media for storing and publishing the statistics it produces, for example on-line databases, diskettes, and CDROM disks. By and large, the documentation templet for published statistics presented in figure 2.3 above will also be adequate for describing metadata, which are stored and published in electronical form. However, figure 2.4 presents an alternative documentation templet for electronically stored macrodata sets. It takes into account the needs for providing detailed and formalized descriptions of the physical and technical aspects of stored macrodata in addition to the more contents-oriented description already covered by the templet in figure 2.3.

Exactly which documentation items should - in a practical situation - be contained in parts 4 and 5 of the macrodata documentation templet will be dependent upon the hardware and software used for storage and retrieval. Here we have assumed that the macrodata are stored in some kind of flat files or relational database. If some completely different data organization method is chosen, the documentation templet would have to be modified accordingly.

0 Data set 0.1 Identification 0.2 Short verbal description 0.3 Data-providing surveys and systems 0.4 Responsibility: organization, unit, person 0.5 Physical location	**1 Contents** 1.1 Verbal description 1.2 Object types and populations: names, definitions, relations 1.3 Subsets of objects accounted for 1.4 Major variable groups and variables 1.5 Structured overview of estimated statistical characteristics 1.6 Frequency, if applicable 1.7 Comparability in time and with other data
2 Errors and accuracy 2.1 Sampling 2.2 Non-response 2.3 Measurement/observation 2.4 Frame coverage 2.5 Processing 2.6 Model assumptions 2.7 Summary description of the total error	**3 Availability** 3.1 Physical location of data 3.2 Secrecy conditions 3.3 Retrieval procedures, including price information 3.4 Related documentation 3.5 Related statistics and analyses of the data
4 Physical characteristics 4.1 Storage form and physical organization 4.2 Data volumes: number of records, record and block sizes, etc 4.3 Key data: primary key, foreign keys, sort key	**5 Record contents and layout** *For each variable/field:* 5.1 Variable/field name 5.2 Short verbal description 5.3 Information source, measurement instrument (including indication of derived variable) 5.4 Definition (informal and/or formal) 5.5 Data type 5.6 Location within record 5.7 Valid values, explicitly stated or indicated by reference to code register
6 Log-book 6.1 Major changes in comparison with earlier versions of the data set, if applicable 6.2 Production events and costs 6.3 Retrieval events and revenues	

Figure 2.4. *Documentation templet for a stored macrodata set.*

2.4 Documentation of observation registers

Observation registers, containing observed and/or derived microdata, are - beside collections of statistics/macrodata - the other important type of information output from statistical surveys. More and more competent users of statistics demand access to microdata, for their own analyses, in their own computer environments. Statistical offices are responding to such demands by preparing files of **anonymized microdata**, for example so-called **public files**.

An external user who is about to (re)use the microdata in an observation register may not be in a position where he or she has access to the staff in the statistical office, who once (maybe many years earlier) produced the data. Thus the observation register will have to be accompanied by an appropriate documentation. Actually the internal staff of the statistical office will often have similar needs, since they, too, are using and reusing microdata, which others have produced.

2.4.1 A survey oriented documentation templet

Figure 2.5 shows a documentation templet for observation registers. The documentation templet focuses a lot on the survey underlying the observation register. However, it is believed that in order to understand properly the meaning and quality of the microdata stored in an observation register, a user will have to know quite a lot about the survey that has produced the data:

(i) what the survey aimed at measuring - the **survey contents**;
(ii) how the survey was designed - the **survey plan**; and
(iii) what actually happened, when the survey was carried out, for example which **errors and deviations** from the survey plan that occurred;

These important aspects of the survey behind the observation register are covered by parts 1, 2, and 3 of the documentation templet in figure 2.5. Items 2.5 and 3.5 contain the descriptions, which more directly concern the observation register as such, as well as the data sets which constitute the physical representation of the observation register.

Part 4 of the documentation templet contains information about the statistical processing of the observation register, which was carried out during the survey. It describes the observation models and the estimation models, which were used when the observation register was processed for the first time (at survey production time), which estimations were actually made then, and how the results were reported and analyzed. This information could be claimed to be redundant for a reuser of the observation register, who is actually free to make his/her own judgements and choices concerning models, estimations, analyses, and presentation. In fact, a suitable criterion for a good observation register documentation is precisely this - that it should enable a (re)user of the data to make these decisions independently of the decisions made by the original producer of the data. On the other hand, it is often very practical for a (re)user of an observation register to know which decisions concerning models, estimations, analysis, and presentation that were made at survey production time. If nothing else, it could speed up the (re)user's work considerably.

0 Documentation structure etc 0.0 Documentation templet 0.1 Survey described 0.1.1 Survey identifier and responsible person 0.1.2 System identifier and responsible person 0.1.3 Organizational unit responsible 0.2 Documentation modules and subsystems 0.3 Survey outputs 0.4 Other relevant documentation	**1 Survey contents** 1.1 Universe of interest and target universe, verbal description 1.2 Target universe, formal description 1.2.1 Target objects 1.2.1.1 Verbal description 1.2.1.2 Object graph 1.2.2 Target populations 1.2.3 Target variables 1.3 Statistical outputs from the survey 1.3.1 Semantical structures 1.3.2 Distribution forms
2 Survey plan 2.1 Frame procedure and observation objects 2.1.1 Overview 2.1.2 Frame and its links to objects 2.2 Sampling procedure (if applicable) 2.3 Overcoverage and undercoverage 2.4 Data collection procedure 2.4.1 Information sources 2.4.2 Measurement instruments 2.4.3 Interruptions 2.4.4 Substitutions 2.5 Planned observation register 2.5.1 Overview 2.5.2 Object types 2.5.3 Object graph 2.5.4 Object/variable-matrixes	**3 Data collection and data preparation** 3.1 Sampling (if applicable) 3.2 Data collection 3.2.1 Communication with the information source 3.2.2 Measurement 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 3.4 Production of 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 Frequency counts of overcoverage, responses, non-responses, etc 3.5 Archiving and dissemination of microdata 3.5.1 Overview 3.5.2 Data set descriptions
4 Statistical processing and reporting 4.1 Observation models 4.1.1 Sampling 4.1.2 Non-response 4.1.3 Measurement/observation 4.1.4 Frame coverage 4.1.5 Total model 4.2 Estimation models 4.3 Estimations 4.3.1 Point estimations 4.3.2 Estimations of sampling errors 4.3.3 Estimations of other quality characteristics 4.4 Analyses 4.5 Presentation and dissemination 4.5.1 Reporting through printed outputs 4.5.2 Electronical dissemination including databases	**5 -**
6 Log-book 6.1 Major changes affecting comparability 6.2 Production events and costs 6.3 Retrieval events and costs	

Figure 2.5. *Documentation templet for an observation register.*

The emptiness of part 5 in the documentation templet requires an explanation. As will be explained in chapter 3, it is desirable that observation register documentations can be produced as automatically as possible from production system documentations. As we shall see the proposed documentation templet for production systems has a structure and contents which are very similar to the documentation templet for observation registers that we are discussing now. However, the documentation templet for production systems has a need for more detailed information about all processes in the production system, and this information is contained in part 5 of the production system documentation (cf figure 2.7). It is not needed by (re)users of observation registers.

2.4.2 A data set oriented documentation templet

Figure 2.6 shows an alternative documentation templet for observation registers, where the metainformation is more centred around the microdata set(s) representing the observation register. Parts 4 and 5 explicitly deal with physical characteristics of the stored microdata set and its records, but it should be noted that the metainformation provided for each variable/field includes some of the contents-oriented metainformation to be provided under item 2.5.4 in the corresponding survey oriented documentation templet in figure 2.5.

It could also be noted that the data set oriented documentation templet in figure 2.6 has inherited some of the "quality declaration flavour" of the documentation templets for published statistics and stored macrodata sets in figures 2.3 and 2.4. Parts 1, 2, and 3 cover the three major aspects of a quality declaration: relevance, accuracy, and availability.

Like for the corresponding macrodata documentation templet (figure 2.4), the exact contents of parts 4 and 5 in the microdata documentation templet in figure 2.6 will depend upon the hardware and software used for storage and retrieval. Once again we have assumed that the data are stored in some kind of flat files or relational database. If some completely different data organization method is chosen, the documentation templet would have to be modified accordingly.

0 Data set 0.1 Identification 0.2 Short verbal description 0.3 Data-providing survey or information system 0.4 Responsibility: organization, unit, person 0.5 Physical location	**1 Contents** 1.1 Verbal description 1.2 Object types and populations: names, definitions, relations 1.3 Important subsets of objects: strata, subdomains of interest, etc 1.4 Sample characteristics, including weights 1.5 Major variable groups and variables 1.6 Information sources and instruments for observation/measurement 1.7 Frequency, if applicable 1.8 Comparability in time and with other data
2 Errors and accuracy 2.1 Sampling 2.2 Non-response 2.3 Measurement/observation 2.4 Frame coverage 2.5 Processing 2.6 Model assumptions 2.7 Summary description of the total error	**3 Availability** 3.1 Physical location of data 3.2 Secrecy conditions 3.3 Retrieval procedures, including price information 3.4 Other relevant documentation 3.5 Available statistics and analyses of the data
4 Physical characteristics 4.1 Storage form and physical organization 4.2 Data volumes: number of records, record and block sizes, etc 4.3 Key data: primary key, foreign keys, sort key	**5 Record contents and layout** *For each variable/field:* 5.1 Variable/field name 5.2 Short verbal description 5.3 Information source, measurement instrument (including indication of derived variable) 5.4 Definition (informal and/or formal) 5.5 Data type 5.6 Location within record 5.7 Valid values, explicitly stated or indicated by reference to code register 5.8 Frequencies of the respective valid values
6 Log-book 6.1 Major changes in comparison with earlier versions of the data set, if applicable 6.2 Production events and costs 6.3 Retrieval events and revenues	

Figure 2.6. *Documentation templet for a stored microdata set.*

2.5 Documentation of survey production systems

The metainformation specified by the documentation templets for published statistics (figure 2.3), stored macrodata (figure 2.4), observation registers (figure 2.5), and stored microdata (figure 2.5) should be sufficient to satisfy the needs of users of statistics and statistical microdata. In order to cover the metainformation needs of statistics producers, a more detailed and complete documentation of all the processes and data sets of a statistical production system will sometimes be needed. In particular, the staff who are responsible for the operation and maintenance of computerized data processing routines will need a more detailed and complete documentation of the data processing system, including processes and data sets of an auxiliary nature. Documentation of auxiliary processes and data is usually not needed by the users for their understanding and use of the information outputs from the system (observation registers and statistics).

Figure 2.7 presents a documentation templet for a survey production system. The main difference between this documentation templet and the documentation templet for observation registers in figure 2.5 is that it contains a part 5, giving full details about the data processing system. The documentation templet suggests that this documentation is structured into three major parts:

(i) survey preparation (including sampling); documentation item 5.1;
(ii) data collection and data preparation; documentation item 5.2;
(iii) estimations, analyses, and reporting; documentation item 5.3.

These three parts correspond to the three major phases in the operation of a statistical survey.

Other differences between the documentation templets for observation registers and production systems are mainly caused by the circumstance that a production system documentation should, in principle, be coninuously updated, so that it always shows the *current status* of the production system, whereas an observation register documentation should describe the *historical status* of an observation register (and the production system behind it) *at a certatin time*, the time when the observation register was produced.

0 Documentation structure etc 0.0 Documentation templet 0.1 Survey described 0.1.1 Survey identifier and responsible person 0.1.2 System identifier and responsible person 0.1.3 Organizational unit responsible 0.2 Documentation modules and subsystems 0.3 Survey outputs 0.4 Other relevant documentation	**1 Survey contents** 1.1 Universe of interest and target universe, verbal description 1.2 Target universe, formal description 1.2.1 Target objects 1.2.1.1 Verbal description 1.2.1.2 Object graph 1.2.2 Target populations 1.2.3 Target variables 1.3 Statistical outputs from the survey 1.3.1 Semantical structures 1.3.2 Distribution forms
2 Survey plan 2.1 Frame procedure and observation objects 2.1.1 Overview 2.1.2 Frame and its links to objects 2.2 Sampling procedure (if applicable) 2.3 Overcoverage and undercoverage 2.4 Data collection procedure 2.4.1 Information sources 2.4.2 Measurement instruments 2.4.3 Interruptions 2.4.4 Substitutions 2.5 Planned observation register 2.5.1 Overview 2.5.2 Object types 2.5.3 Object graph 2.5.4 Object/variable-matrixes	**3 Data collection and data preparation** 3.1 Sampling (if applicable) 3.2 Data collection 3.2.1 Communication with the information source 3.2.2 Measurement 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 3.4 Production of 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* 3.5 Archiving and dissemination of microdata 3.5.1 Overview 3.5.2 Data set descriptions
4 Statistical processing and reporting 4.1 Observation models 4.1.1 Sampling 4.1.2 Non-response 4.1.3 Measurement/observation 4.1.4 Frame coverage 4.1.5 Total model 4.2 Estimation models 4.3 Estimation algorithms 4.3.1 Point estimations 4.3.2 Estimations of sampling errors 4.3.3 Estimations of other quality characteristics 4.4 Analyses 4.5 Presentation and dissemination *4.5.1 Reporting through printed outputs* 4.5.2 Electronical dissemination 4.5.3 Dissemination through databases	**5 Data processing system** 5.0 System overview 5.0.1 Verbal description 5.0.2 System flow 5.1 Survey preparation (including sampling) 5.1.1 Overview 5.1.1.1 Verbal description 5.1.1.2 System flow 5.1.2 Component descriptions 5.1.2.1 Data sets 5.1.2.2 Processes 5.1.2.3 Other components 5.2 Data collection and data preparation 5.2.1 Overview 5.2.1.1 Verbal description 5.2.1.2 System flow *5.2.2 Component descriptions* 5.2.2.1 Data sets 5.2.2.2 Processes 5.2.2.3 Other components 5.3 Estimations, analyses, and reporting 5.3.1 Overview 5.3.1.1 Verbal description 5.3.1.2 System flow 5.3.2 Component descriptions 5.3.2.1 Data sets 5.3.2.2 Processes 5.3.2.3 Other components
6 Log-book	

Figure 2.7. *Documentation templet for a survey production system.*

2.6 Register of survey-related events

All documentation templets presented so far have contained a final part called **log-book**. The log-book may have all or some of the following purposes:

(i) to keep track of design changes as they occur in a survey production system over time, so as to make it possible to reconstruct the properties of the production system and its information outputs (observation registers and statistics) at some earlier stage;

(ii) to give an overview of important design changes between different time versions of information outputs (observation registers and statistics) from the production system;

(iii) to collect production-oriented statistics, so as to facilitate improvements in the production efficiency;

(iv) to collect retrieval-oriented statistics, so as to facilitate improvements in user satisfaction;

(v) to support cost accounting, pricing decisions, and invoicing routines.

A **register of survey-related events** could be seen as a conceptualization of a metainformation system tool comprising the above-mentioned tasks.

2.7 Retrieval-related knowledge

2.7.1 Retrieval system documentation

Retrieval systems are designed and operated from a more pronounced user perspective than traditional survey production system. However, it is felt that the considerations and proposals concerning documentation templets, etc, that have been put forward in these *Guidelines*, can relatively easily be modified so as to be useful for retrieval systems as well as for production systems. This is a conjecture, which will have to be investigated further.

2.7.2 Reference data: tables of contents, indexes, etc

If we look at the specification templet in figure 2.1, there are several tasks which require global searches through large amounts of data and metadata, emanating from many different surveys and information systems. Items 1.1 and 3.1 contain example of such tasks. Beside being global, these searches also have the feature in common that the persons initiating the searches may have very different frames of references and different ways of thinking and associating, which will heavily influence their ways of putting questions and their methods of performing searches.

A conclusion from the observations just made is that a statistical metainformation system must contain a substantial amount of **reference data** that will provide a basis for alternative, flexible, and user-friendly ways of searching and associating statistical data and metadata in a large, global statistical information system. Some common forms of reference data are **tables of contents** and **indexes** of different kinds. Another important type of tool is the **statistical thesaurus** (cf section 2.8 below).

2.7.3 Register of retrieval-related events

A register of retrieval-related events will have similar tasks as a register of survey-related events (cf section 2.6 above). It will be a kind of log-book, focusing on the following tasks:

(i) collecting retrieval-oriented statistics, so as to facilitate improvements in user satisfaction;

(ii) providing feed-back to the underlying surveys by informing the producers responsible about retrieval frequencies, user satisfaction, user complaints and demands for other input, etc;

(iii) supporting cost accounting, pricing decisions, and invoicing routines;

(iv) keeping track of important design changes in the retrieval system as well as in underlying surveys.

2.8 General knowledge

General knowledge comprises metainformation, which is not directly related to specific, individual surveys. A comprehensive statistical metainformation system could contain the following major categories of general knowledge:

(i) handbook knowledge;
(ii) encyclopedical knowledge;
(iii) standards (and metainformation related to standards);
(iv) software and software-related knowledge.

2.8.1 Handbook knowledge

Handbook knowledge is knowledge about how a major task could be accomplished from the beginning to the end. An important piece of handbook knowledge to be contained in a comprehensive statistical metainformation system would be a

- ***handbook in the design of statistical surveys and information systems***

Such a handbook would tell how to design a statistical survey, or a statistical information system, from the beginning to the end, covering (for a statistical survey)

(i) survey purpose, contents, and outputs;

(ii) the survey plan, including
- frame procedure;
- sampling procedure (if applicable);
- overcoverage and undercoverage;
- data collection procedure;
- data preparation;
- observation register;

(iii) statistical processing and reporting, including
- observation models;
- estimation models and estimation algorithms;
- analysis, presentation, and dissemination.

It should be noted that the structure of the earlier presented documentation templets for survey production systems and observation registers concide quite well with the structure proposed here for handbook knowledge concerning the design of a statistical survey.

Another relevant piece of handbook knowledge to be contained in a comprehensive statistical metainformation system would be a

- **handbook in project evaluation and project management**

This knowledge would correspond to some metainformation needs of designers and managers that were listed in the specification templet in figure 2.1.

2.8.2 Encyclopedical knowledge

The development of a complete handbook is a rather ambitious task. An **encyclopedia** could be a good complement or even an alternative to a handbook. It would contain short articles describing, for example, a certain statistics production method.

Other types of encyclopedical knowledge are dictionaries and thesauri. A **dictionary of statistical terminology** would describe important terms used by designer of statistical surveys and information system.

A **statistical thesaurus** could be an important tool for facilitating searches in large statistical databases. A statistical thesaurus explains the meaning of statistical characteristics covered by a statistical information system, but it does more than this. It relates the terms used in naming the statistical characteristics to other terms, for example:

- (more or less) **synonymous terms**;
- **broader terms**;
- **narrower terms**;

Thus a statistical thesaurus facilitates **associative searches**, and it helps to overcome situations, where different users (and producers) of statistical data are using different terms when referring to (more or less) the same characteristics and phenomena.

A statistical thesaurus could also facilitate the **integration of statistics** from different countries with different languages. For example, it could help to achieve consistent translations of table titles etc.

2.8.3 Standards

Statistical standards could be subdivided into **contents-oriented** and **representation-oriented** standards. To some extent they are closely related to each other. For example, a code list of standard values of a certain variable or classification would often contain a description of a preferred representation format as well as contents-oriented definitions of the variable and the respective values.

2.8.4 Software and software-related knowledge

Statistical software may in itself contain important knowledge about statistical algorithms and procedures. Such information may be important to make available in a statistical metainformation system, together with information about how to use the software.

3 Implementation strategies for statistical metainformation systems

In section 1 we identified the main users and usages of metainformation involved in the different phases of the life cycle of a statistical survey or a statistical information system. In section 2 we then identified some major categories of metainformation and metainformation-related system functions that are needed to satisfy the needs by the users and usages identified in section 1.

In this section we shall provide some guidelines concerning how to construct and implement a well-functioning metainformation system that could provide the metainformation and the functionality described in section 2, or at least some parts thereof, which are given priority by the metainformation users.

3.1 Short-term priorities vs long-term completeness and rationality

It is usually overambitious and unrealistical to plan the immediate development of a "complete" metainformation system, satisfying "all" needs for metainformation and metainformation-related functionality. This goes more or less without saying. Moreover, this thesis is supported by several disappointing practical experiences in the past.

On the other hand, it is equally dangerous to plan the development of a metainformation system with an extreme focus on a few particularly important needs, putting on blinkers to related needs and prerequisites. For example, while it is natural and highly commendable for a modern statistical office to give priority to the metainformation needs of statistics users, it would be stupid to neglect interdependencies between such a task and other functions of a more complete metainformation system infrastructure. It would be like the city child asking his mother, why cows are needed, when we can buy milk in the stores, or the statistics-hater criticizing government for spending a lot of money on producing unemployment statistics, when anyone can read unemployment figures in the newspapers every day.

In this first version of the *Guidelines* we *shall* assume, as just suggested, that it *is* the needs of the statistics users that should be given top priority, realizing at the same time that the metainformation needed by the statistics users has to be produced in an economical way and in such a way that the metainformation and functionality produced will be of adequate quality.

We advocate the approach that a statistical office (or some other kind of statistical organization) should

(i) *first* define a **target architecture** for its entirety of metainformation systems - the **metainformation infrastructure** of the statistical office - as it should be organized in a long-term perspective in order to satisfy foreseeable needs for metainformation and metainformation-related functions in an efficient way; and

(ii) *then* implement the metainformation infrastructure **incrementally**, step by step, starting with

(a) the subsystems, which are most urgently needed by the statistics users; and

(b) subsystems, which are needed for an efficient operation of the subsystems most urgently needed by the statistics users.

With this approach it is natural to focus on both the production and the usage aspects of the operation phase of the life cycle of a statistical information system. A somewhat lower priority will be given to the metainformation needs occurring in the design and management phases of the life cycle. Obviously, in accordance with what we have just stated in (ii)(a) above, priority should be given to the metainformation subsystems needed for the user-oriented aspects of the operation of a statistical information system. However, since certain production-oriented subsystems are essential for the efficient functioning of the user-oriented subsystems, these production-oriented subsystems should also be given priority - in accordance with what was stated under (ii)(b) above. We shall return return later to the question, which these production-oriented subsystems are that are essential for the efficient functioning of the user-oriented subsystem.

An additional reason for giving somewhat lower priority to metainformation subsystems, which are oriented to the needs of designers and managers of statistical information systems, is that these needs (and metainformation subsystems responding to these needs) are still less explored and less formalized than those of users and producers of statistics. Moreover, one may relatively safely assume that the metainformation needed by users and producers of statistics will anyhow turn out to be a natural and necessary basis for the metainformation holdings and metainformation-related functions needed by designers and managers.

3.2 Capture metadata early - and never twice

In order to produce the metainformation needed by a statistical office and its users in a rational and economical way, one should

(i) capture metadata at the time and place where they are most naturally born, that is, when and where they occur for the first time;

(ii) avoid to use scarce human resources for manually copying metadata, which already exist, or which can be automatically produced from existing metadata by means of formal transformations, which can be computerized.

In order to be able to follow these simple but extremely important rules, and in order to draw all consequences of them, we must make a careful analysis of all processes in a statistical office, which are potential users and/or producers of metadata, and we must study how these processes are (or can be designed to be) linked to each other in a such a way that the two rules stated above can be obeyed. We have a good starting-point for the analysis in the previous chapters of these *Guidelines.*

3.3 The data/metadata symbiosis

When analyzing how the metadata flows of a statistical office should be designed, we shall soon find that the metadata flows need to be very closely linked to the data flows. In fact the metadata flows need to live in a kind of symbiosis with the data flows. This is not very surprising. As a matter of fact, before statistics production was computerized, data and metadata were always handled in an integrated way througout the production process, from the questionnaires with their questions (metadata) and entered data values, to the published tables with their figures (data) and labels, texts, and footnotes (metadata) describing the meaning of the data.

Figure 3.1 gives an overview the most important data and metadata flows in a statistical office and their natural relationships to each other. At the same time it indicates the most important data and metadata holdings and their natural positions in the data/metadata flow.

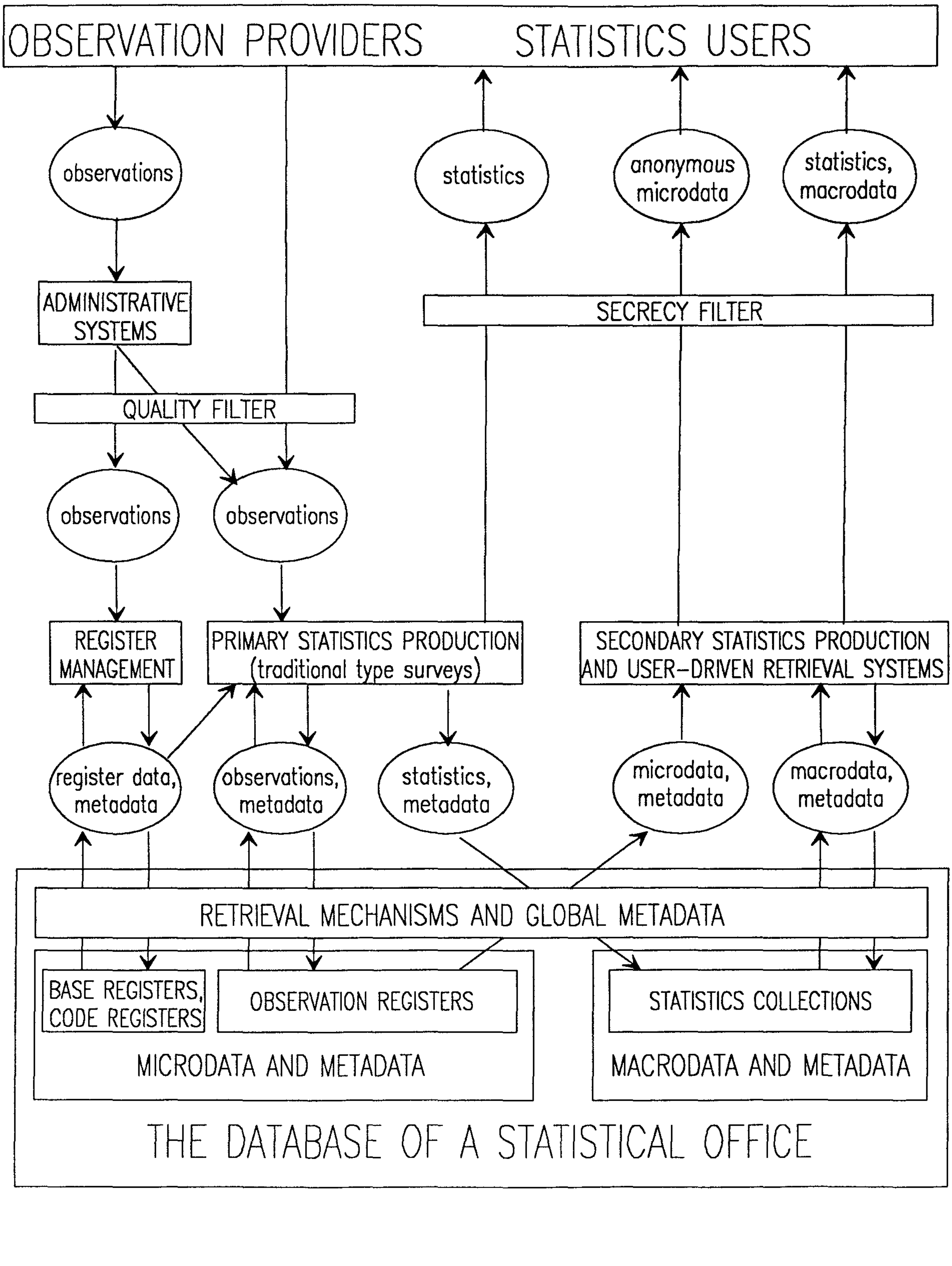


Figure 3.1. *The statistical information system of a statistical office.*

The general direction of the data/metadata flow in figure 3.1 is

- *from* observation providers,
- *through* administrative systems and/or primary statistics production systems (surveys),
- *into* observation registers and statistics collections,
- *through* secondary statistics production systems and/or user-driven retrieval systems,
- *to* statistics users.

As we shall discuss later, there are important **feed-back loops** in the flow as well, in particular if we consider design-oriented and management-oriented data/metadata flows in additon to the operation-oriented flows shown in figure 3.1.

The model in figure 3.1 makes a distinction between on the one hand

(i) **primary statistics production**; and, on the other hand
(ii) **secondary statistics production** and **user-driven retrieval systems**.

The primary statistics production systems are more or less equivalent with what we have earlier called the traditional statistical surveys, and the user-driven retrieval systems are equivalent with the retrieval systems that we have discussed.

The secondary statistics production systems - systems like the production system of the *system of national accounts* of a country - are sometimes a little difficult to categorize. They distinguish themselves from the primary statistics production systems by obtaining all (or virtually all) their input data (macrodata and/or microdata) from other statistics production systems, that is, they do not themselves collect direct observations. On the other hand, the secondary production systems distinguish themselves from the more genuine, user-driven retrieval systems by being essentially producer-driven, and by focusing on some very particular user and purpose.

Figure 3.1 puts **registers** and **register management** into the data/metadata flow model. The registers are subdivided into **base registers**, containing authorized lists of the objects belonging to a certain population, and **code registers**, containing authorized lists of the values belonging to the value set of a certain variable or classification. For many practical purposes the management of a register can be looked upon as the operation and management of a statistical survey. Among other things, the documentation templets proposed earlier in these *Guidelines* for statistical surveys should be useful for registers as well.

3.4 Interdependencies between design, operation, and management phases

It should be noted that figure 3.1 shows the data/metadata flow in the way which is most relevant from an operation point of view. It does not illustrate the production and use of data and metadata in design and management processes. However, it is important that

(i) the design-oriented data/metadata flows;
(ii) the operation-oriented data/metadata flows; and
(iii) the management-oriented data/metadata flows;

are analyzed together, since there are many important interdependences between them. For example, many metadata are naturally generated and captured as the result of design processes and design decisions, and many of these metadata (for example names, definitions, and storage formats of different statistical characteristics and their components) are later needed (at operation time) by producers and users of statistical data, as well as by software products supporting the tasks of users and producers.

Figure 3.2 makes an attempt to illustrate some of the metadata interdependencies between different phases in the life cycle of a statistical information system. It also suggests certain interdependencies between different kinds of statistical information systems, primarily production systems and retrieval systems, but also auxiliary systems like registers. In the figure "System 1" has been indicated as a production system and "System 2" as a retrieval system, but this is of course an arbitrary choice. Among the supposedly n different systems in the whole complex, there may be any number of production systems, retrieval systems, and auxiliary systems like registers, and many of them will have data/metadata-interdependences between them.

Figure 3.2 shows an important feed-back loop from the n local metadatabases to the common global metadatabase. The local databases contain detailed knowledge concerning specific individual systems and their specific statistical data (observation registers and statistics). The system-specific local metadata have to be processed (preferably as automatically as possible) in order to create extracts and summaries, which can be managed in the global metadatabase, from which it can be easily retrieved by local systems as well as external systems. In order to make the retrieval of global metadata as efficient and user-friendly as possible, the extracts and summaries have to be further processed (once again as automatically as possible) in order to create and maintain reference data like tables of contents and indexes.

The generation of extracts, summaries, and reference data is one part of the knowledge acquisition process for the global metadatabase. Another part is the acquisition of general knowledge: handbooks, encyclopedia, thesauri, standards, software, etc. The acquisition of general knowledge can be performed rather independently of the feed-back loop just described. However, there is a potential for creating an "intelligent" **inductive learning loop** from the local and global specific knowledge to the global general knowledge. At the present state of the art this inductive learning loop will be highly dependent on human efforts, but artificial intelligence may contribute increasingly in the future.

It is difficult to represent interdependencies between local and global metadata, between different life cycle phases, and between different statistical information systems in one and the same model, illustrated by one and the same figure. If one wants to give a more precise description of the metadata flows and their interdependencis, one has to concentrate on one aspect, or a small number of aspects, at a time. In comparison with figure 3.2, figure 3.3 shows metadata flows for one statistical information system only, a survey production system. However, for this single system, figure 3.3 gives a more precise description of the metadata flows for each one of three major phases of the life cycle of the system: design, operation, and evaluation.

Figure 3.3 indicates how metadata holdings and metadata flows could be organized so as to satisfy some of the requirements given by the specification templet in figure 2.1. Thus the metadata flows entering and leaving the design process in figure 3.3 correspond to item 3.1 in figure 2.1, the flows to and from the operation process correspond to items 2.1 and 2.2, and the flows linked to the evaluation process correspond to items 3.2 and 4.1 in the specification templet.

Still the model in figure 3.3 may be felt to contain too much information at a time. In figure 3.4 the same information has been broken down into three submodels, each one of them focusing on one phase of the life cycle of the survey production system.

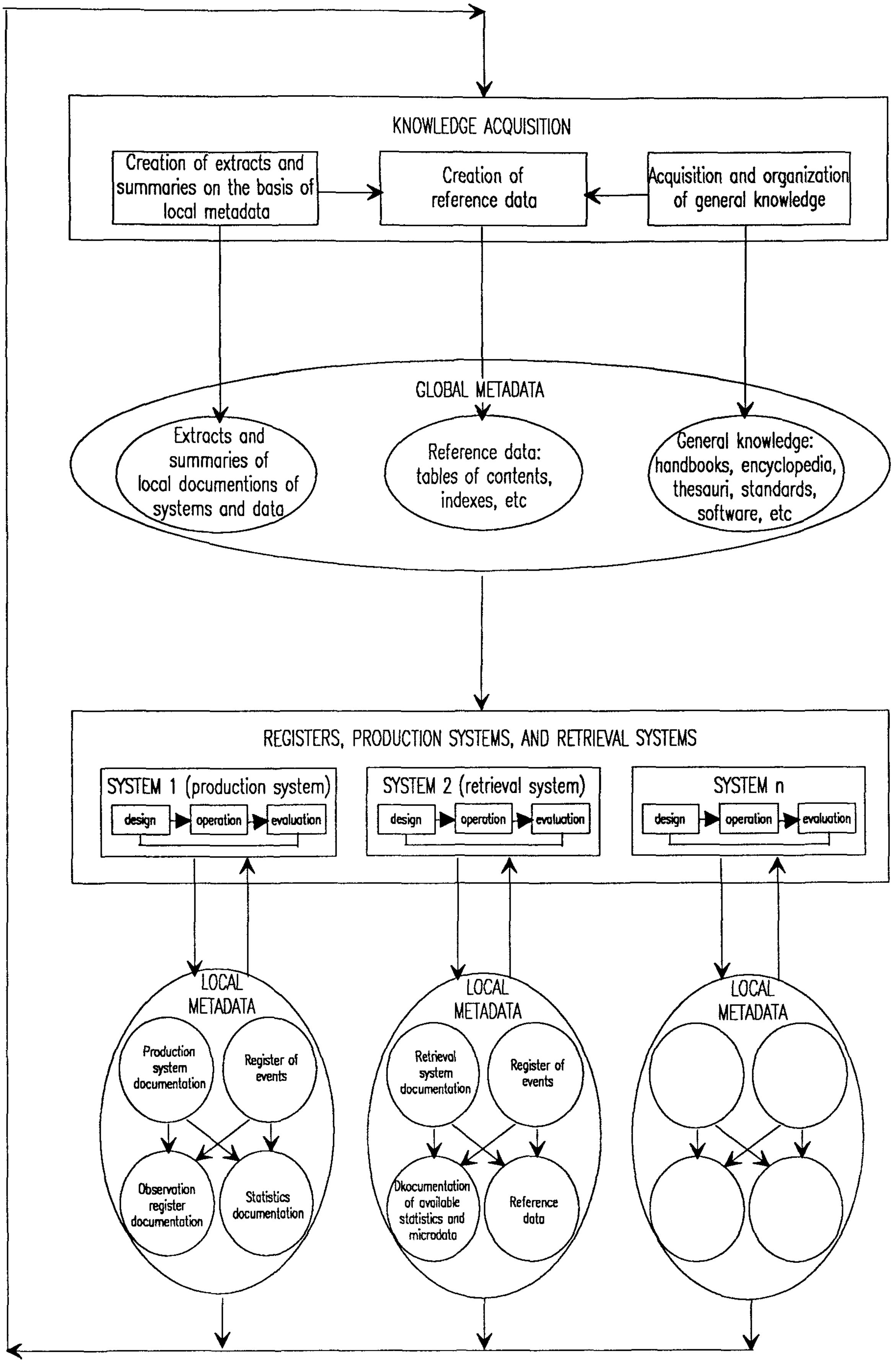

KNOWLEDGE ACQUISITION
Creation of extracts and summaries on the basis of local metadata
Creation of reference data
Acquisition and organization of general knowledge
GLOBAL METADATA
Extracts and summaries of local documentions of systems and data
Reference data: tables of contents, indexes, etc
General knowledge: handbooks, encyclopedia, thesauri, standards, software, etc
REGISTERS, PRODUCTION SYSTEMS, AND RETRIEVAL SYSTEMS
SYSTEM 1 (production system)
design
operation
evaluation
SYSTEM 2 (retrieval system)
design
operation
evaluation
SYSTEM n
design
operation
evaluation
LOCAL METADATA
Production system documentation
Register of events
Observation register documentation
Statistics documentation
LOCAL METADATA
Retrieval system documentation
Register of events
Dkocumentation of available statistics and microdata
Reference data
LOCAL METADATA

Figure 3.2. *Metadata flows for a system of different kinds of statistical information systems.*

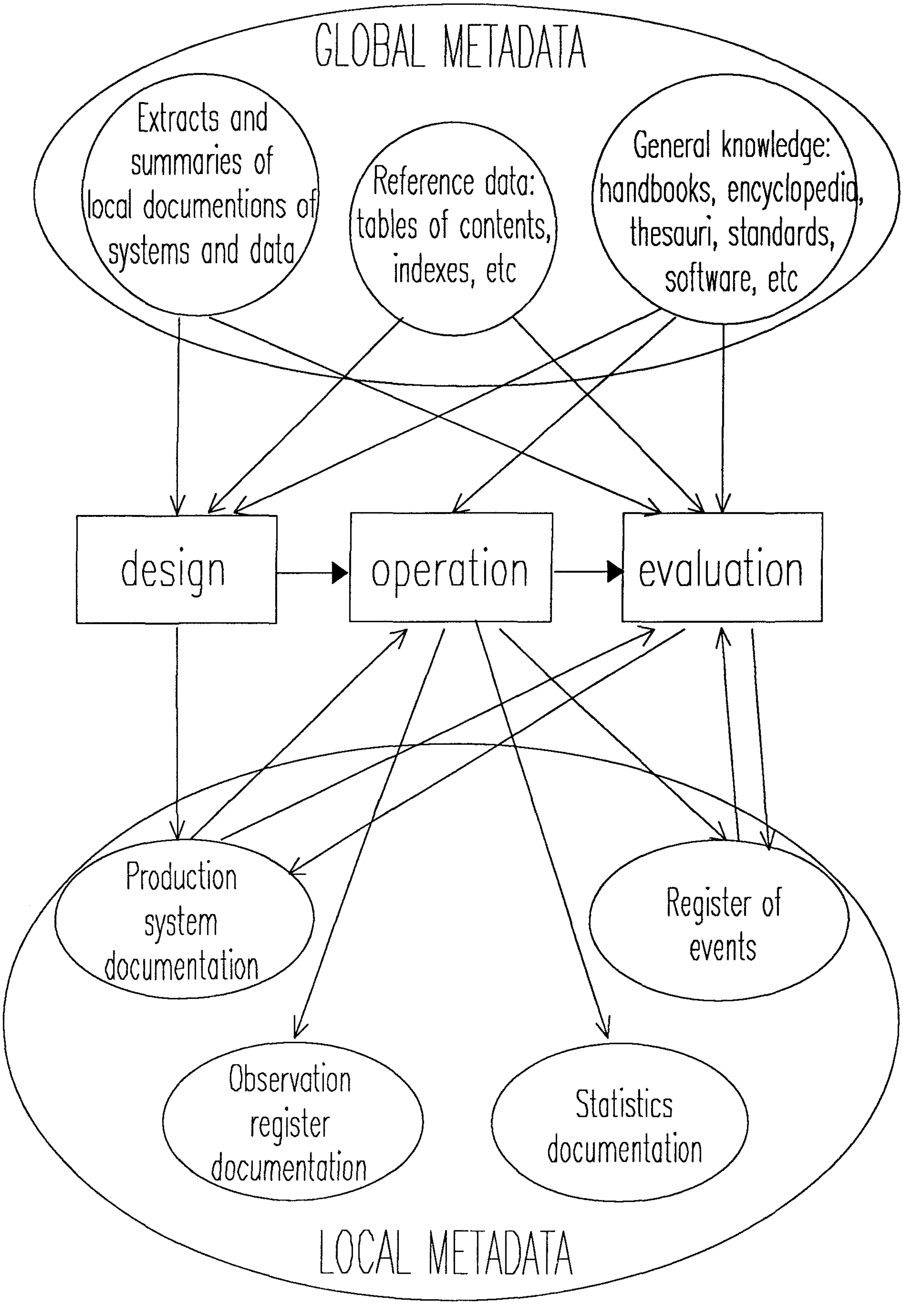

GLOBAL METADATA
Extracts and summaries of local documentions of systems and data
Reference data: tables of contents, indexes, etc
General knowledge: handbooks, encyclopedia, thesauri, standards, software, etc
design
operation
evaluation
Production system documentation
Register of events
Observation register documentation
Statistics documentation
LOCAL METADATA

Figure 3.3. *Metadata flows for the three major phases of the life cycle of a survey production system.*

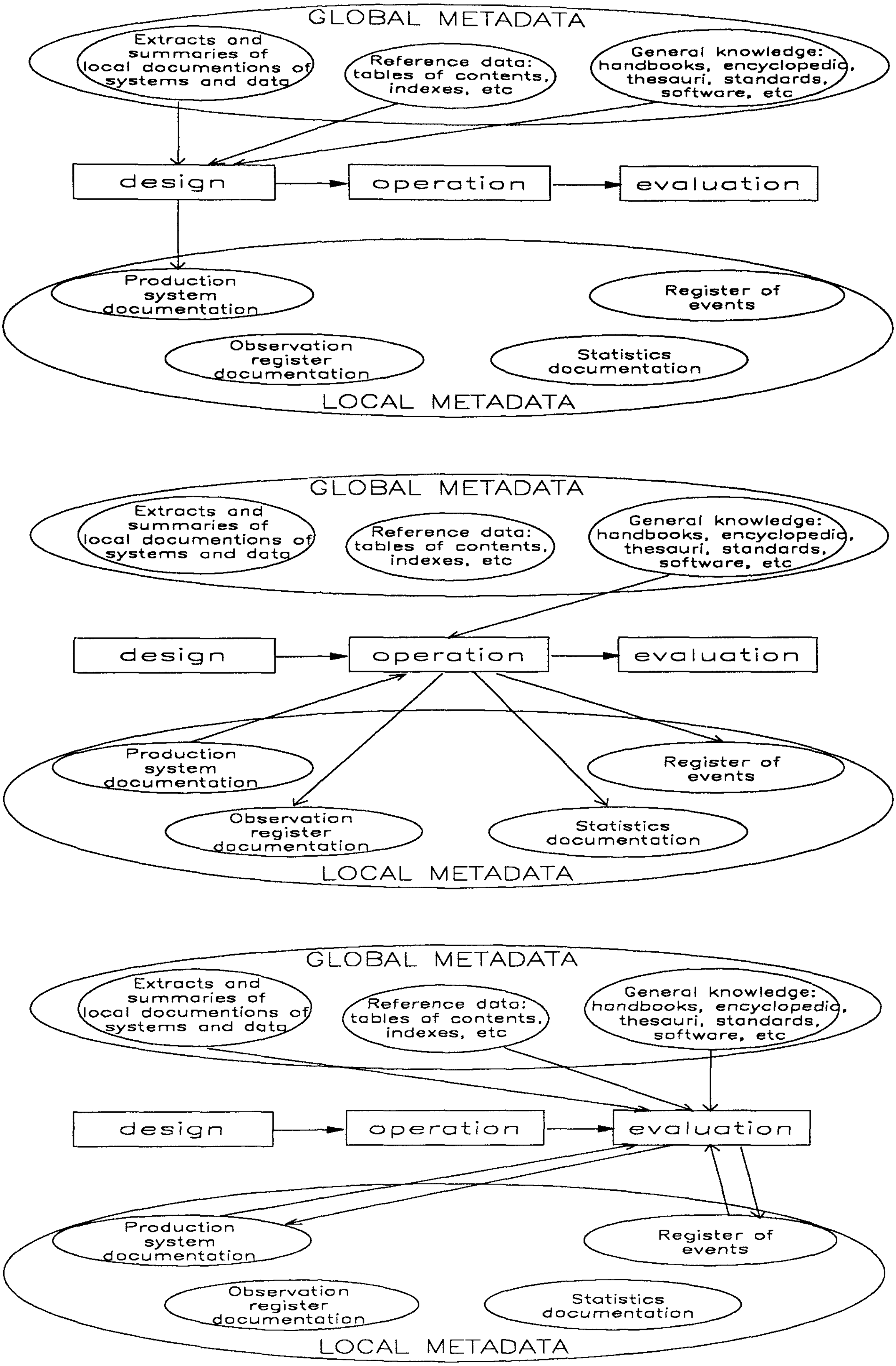

GLOBAL METADATA
Extracts and summaries of local documentions of systems and data
Reference data: tables of contents, indexes, etc
General knowledge: handbooks, encyclopedia, thesauri, standards, software, etc
design
operation
evaluation
Production system documentation
Register of events
Observation register documentation
Statistics documentation
LOCAL METADATA
GLOBAL METADATA
Extracts and summaries of local documentions of systems and data
Reference data: tables of contents, indexes, etc
General knowledge: handbooks, encyclopedia, thesauri, standards, software, etc
design
operation
evaluation
Production system documentation
Register of events
Observation register documentation
Statistics documentation
LOCAL METADATA
GLOBAL METADATA
Extracts and summaries of local documentions of systems and data
Reference data: tables of contents, indexes, etc
General knowledge: handbooks, encyclopedia, thesauri, standards, software, etc
design
operation
evaluation
Production system documentation
Register of events
Observation register documentation
Statistics documentation
LOCAL METADATA

Figure 3.4. *Focusing on the metadata flows for one life cycle phase at a time (cf figure 3.3).*

3.5 Client/server-databases and standardized interfaces

As we have seen, the metainformation infrastructure of a statistical office has to support many complex relationships, for example the relationships between

- production systems and retrieval systems;
- local data/metadata and global metadata;
- different phases in a system life cycle.

The relationships are often complex in the sense indicated in the upper part of figure 3.5, which shows two sets of entities of some kind, which are related to each other in a so-called *"many-to-many"* pattern. For example, one of the two sets could be a set of production systems, and the other one could be a set of retrieval systems.

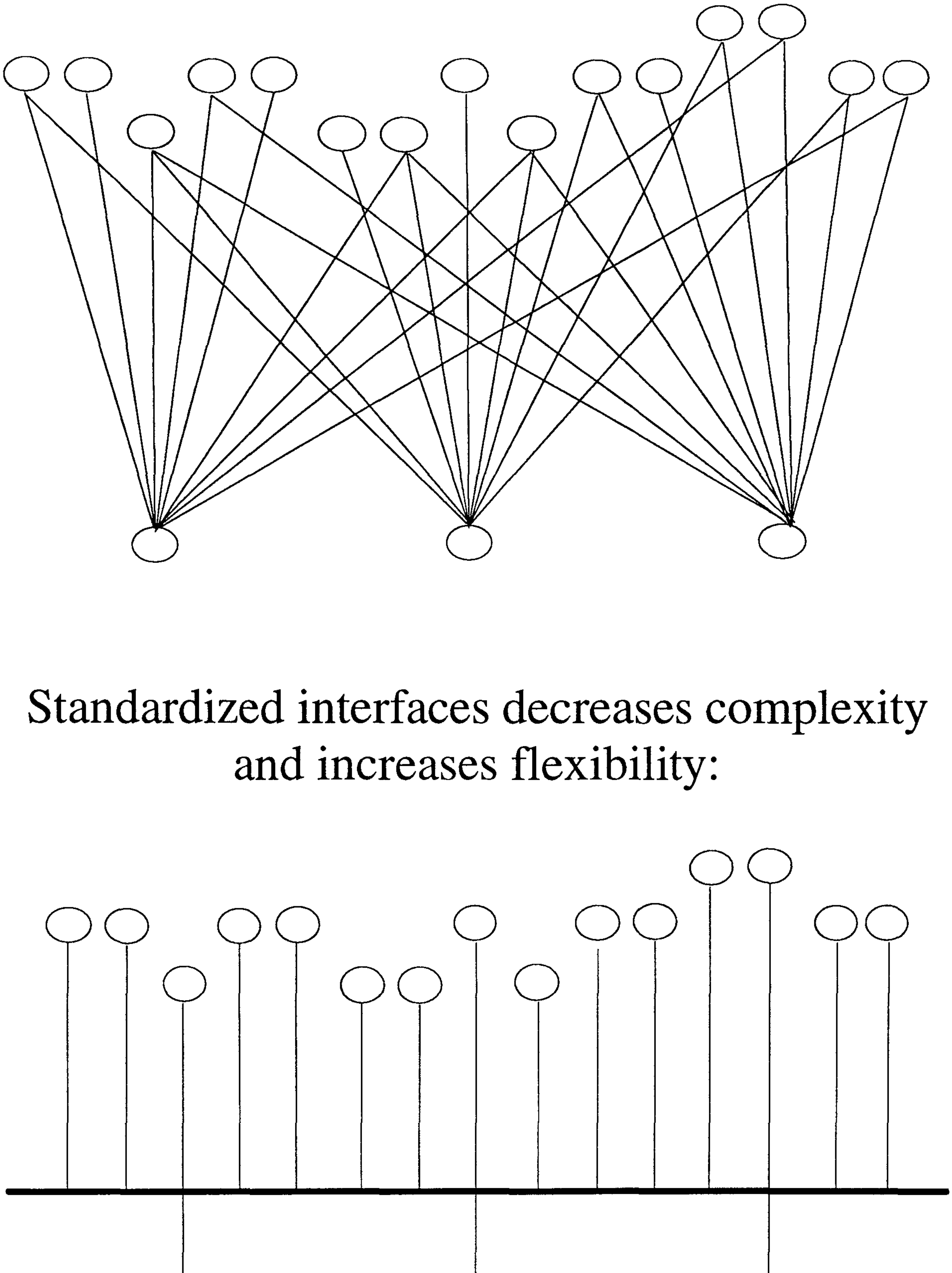
If we solve the communication problem between the two sets of entities in the upper part figure 3.5 by developing a unique interface, or communication channel, between each related pair of entities, the number of interfaces that will have to be developed and maintained will approach a number of the order of magnitude m * n.

If instead we solve the communication problem by developing a standardized, intermediary interface, like indicated in the lower part of figure 3.5, the complexity will be much lower. Instead of communicating directly with each other, the entities in one set will now communicate with the entities in the other set via the standardized, intermediary interface. One has to develop an interface between each one of the entities in anyone of the two sets and the intermediary interface, but the number of such interfaces that will have to be developed and maintain will not be greater than a number in the order of magnitude m + n, that is, a number which is usually considerably lower than m * n. Furthermore, this architecture is much more flexible; if a new entity is added to one of the two sets, all we have to do in order to ensure full communication capabilities between the new entity and all other entities is to develop one communication channel between the new entity and the standardized intermediary interface.

In the models for a statistical data/metadata infrastructure that we have developed and visualized above (figures 3.1 - 3.4) we have actually assumed the existence of some standardized interfaces in the form of interfaces to databases, where the different types of statistical data and metadata are stored:

- observation registers (microdata) and accompanying metadata;
- collections of statistics (macrodata) and accomanying metadata;
- base registers;
- code registers;
- system documentations;
- event registers;
- quality filters;
- secrecy filters;
- reference data;
- general knowledge;

Some of these categories (for example "general knowledge") will have to be subdivided into subcategories, before representation standards can be considered in a meaningful and specific way.



Standardized interfaces decreases complexity and increases flexibility:

Figure 3.5. *Simplifying complex relationships between two sets of entities by introducing an intermediary, uniform interface.*

The development of standards, which are relevant for the above-mentioned purposes, has been initiated by the international statistical community, notably within the EDIFACT framework. On the one hand, it is important the standards become tailored to the special needs of statistical information systems. On the other hand, it is essential that the standards are well in line with commercial standards, established by the software market. In the market situation prevailing now - and probably a few years ahead - it is important that any statistical standards can handle - and make maximum constructive use of - commercial standards like

- relational databases in a client/server-architecture, using an SQL interface;
- Microsoft standards for graphical interfaces (Windows) and software links;
- standards for managing and searching large sets of text data stored on CD-ROM disks, if and when such standards appear.

It is sometimes claimed, for example, that statistical macrodata and accompanying metadata are so special that they cannot be stored and accessed by means of relational software. It is certainly true that relational software alone is not enough to solve all the problems in connection with the management of macrodata and accompanying metadata. However, it should be feasible to develop software for this purpose, which uses relational software (and other established standards and commercial software products) as major components.

3.6 Master plan for the development of a statistical metainformation infrastructure - an outline

There will be no summary at the end of these *Guidelines*. Instead there will be the following outline of a **"master plan"** for how a metainformation infrastructure could actually be developed in practice in a statistical office or some other kind of statistical organization. The presented outline consists of seven steps:

1. Explore the needs for metainformation and metainformation-related functions with potential users in the environment (the statistical office and its users, for example). Use the specification templets suggested in figures 1.3 and 2.1as a structuring tool.

2. Put priorities to the different needs that were identified during the exploration phase.

3. Review the priorities, taking into account that fulfilling some needs maybe instrumental in fulfilling others. For example, production system documentations (cf the documentation templet in figure 2.7) are instrumental for fulfilling many needs, because they contain metadata that are needed, or can easily be transformed into metadata that are needed, for many purposes, including several needs of statistics users, which are typically given high priority.

4. Outline an architecture of a metainformation infrastructure, identifying

(i) **local subsystems**, containing system-specific metadata like production system documentations and documentations of observation registers and statistics produced by specific surveys;

(ii) **global subsystems**, containing

(a) **system-specific metadata**, extracted and summarized from local subsystems;

(b) **general knowledge** (handbooks, encyclopedia, thesauri, standards, etc)

(c) **reference data** to (a) and (b);

(iii) **interfaces and communication channels** for the exchange of metadata (and data) between local and global subsystems - in both directions;

5. Explore suitable standards and software tools for the implementation of the outlined metainformation infrastructure. Consider standards that are already established in the organization, for example through its office information system (word processing system, e-mail, graphical interface etc).

6. Establish a suitable mix of well orchestrated projects. There could be needs for

- a **conceptual project**, developing and marketing a common **professional language** in the organization, so that different categories of specialists, as well as people from different subject matter departments, "speak the same language";

- a project for the development of a **documentation model**, formalized in documentation templets, and using the professional language from the conceptual project;

- one or more projects for developing and/or otherwize acquiring **software tools** that would support acquisition of survey-specific metadata by means of the documentation model;

- one or more **training projects**;

- an effort to get the **survey-specific documentation work** going locally;

- one or more projects for developing **software/data/metadata interfaces** between the most popular software products in the organization and the emerging local and global data/metadatabases, so that people using these software products can actually have their applications automatically fed with appropriate metadata, when they load them with data;

- a small number of projects for developing **strategical global knowledge bases**, like a code register database and a statistical thesaurus;

- a small number of externally oriented projects, aiming at the development attractive **retrieval systems** focusing on the needs of important categories of statistics users.

7. Monitor the selected projects, and **revise** project mix and project plans whenever necessary.

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.

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.

Reports published 1993:

1993:1 Kvalitetsfonden 1991/92 - Projekt, handläggning och utvärdering
(grön) (Roland Friberg)

1993:2 Översyn av Undersökningen av lastbilstransporter i Sverige (UVAV)
(grön) (Bengt Rosén, Mahnaz Zamani)

1993:3 Bortfallsbarometern nr 8 (Mats Bergdahl, Pär Brundell, Jan Hörngren,
(grön) Håkan Lindén, Peter Lundquist, Monica Rennermalm)

Tidigare utgivna **R&D Reports** kan beställas genom Ingvar Andersson, SCB, U/LEDN, 115 81 STOCKHOLM (telefon 08-783 41 47, telefax 08-783 45 99).

R&D Reports can be ordered from Statistics Sweden, att. Ingvar Andersson U/LEDN, S-115 81 STOCKHOLM, SWEDEN (telephone nr 46 08 783 41 47, telefax nr 46 08 783 45 99)