UNITED NATIONS STATISTICAL COMMISSION
and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS
STATISTICAL STANDARDS AND STUDIES – No. 51

INFORMATION SYSTEMS ARCHITECTURE FOR NATIONAL AND INTERNATIONAL STATISTICAL OFFICES

GUIDELINES AND RECOMMENDATIONS

UNITED NATIONS
Geneva, 1999

CONTENTS

THE CONFERENCE OF EUROPEAN STATISTICIANS

The Conference of European Statisticians was set up in 1953 as a continuing body meeting under the auspices of the Economic Commission for Europe and the Statistical Commission of the United Nations. Its objectives are (a) to improve European official statistics and their international comparability having regard to the recommendations of the Statistical Commission of the United Nations, the specialised agencies and other appropriate bodies as necessary; (b) to promote close co-ordination of the statistical activities in Europe of international organisations so as to achieve greater uniformity in concepts and definitions and to reduce to a minimum the burdens on national statistical offices; and (c) to respond to any emerging need for international statistical co-operation arising out of transition, integration and other processes of co-operation both within the ECE region and between the ECE region and other regions. The members of the Conference are the directors of the national statistical offices of the countries participating in the work of the United Nations Economic Commission for Europe. The Conference meets in plenary session once a year and also arranges numerous meetings of specialists on particular statistical subjects.

PREFACE

The methodological material "Information Systems Architecture for National and International Statistical Offices: Guidelines and Recommendations" was reviewed at the Meeting on Management of Statistical Information Technology organised by the United Nations Economic Commission for Europe (UN/ECE) in February 1999 in the framework of the programme of work of the Conference of European Statisticians.

National Statistical Offices of the UN/ECE member countries, the Organisation for Economic Co-operation and Development (OECD), the Food and Agriculture Organisation (FAO), The United Nations Population Fund (UNFPA), and the United Nations Statistics Division participated in this meeting.

The material reflects outcomes of the discussion on the document. The Conference of European Statisticians at its 1999 plenary session decided to publish this material in the Conference's Statistical Standards and Studies Series.

The methodological material was prepared by Professor Bo Sundgren from Statistics Sweden.

SUMMARY

The paper is a methodological report, based upon the author's long-standing experience in designing statistical information systems' (SIS) architectures for national and international statistical organisations. Its goal is to assist statistical offices in designing an efficient information systems architecture under conditions of growing users' demands, increasing international co-operation and constant changes in information technology.

The material provides a comprehensive analysis of the existing types of statistical information systems in national and international statistical offices, and outlines development directions for the future. The paper examines the technical aspects of the proposed SIS architecture, and gives practical recommendations on how to implement a realistic information technology (IT) strategy under conditions of ongoing rapid technological development.

The first chapter explains the basic concepts used in the report: different types of statistical processes and statistical data, statistical applications and infrastructure, and the flow of data and metadata through the survey process.

A statistical information system is composed of subsystems (applications) which collect, process, store, retrieve, analyse and disseminate statistical data. The information systems architecture (ISA) of a statistical office is a common framework within which different subsystems have their respective roles and interact mutually. The paper also analyses some relations between ISA and the organisation architecture of a statistical office, and describes some specifics of an ISA in statistics.

The ISA should reflect the purposes and tasks of the statistical office. One of the reasons for discrepancies between the ISA and the existing organisational architecture of statistical offices could be the conflict between the traditional survey-oriented organisation of statistical offices and the cross-cutting information needs of statistics users. Since surveys are very often navigated by data collections, the organisation of statistical offices is also "input-oriented". This makes it difficult to achieve desirable co-ordination and control across subject-matter areas.

Some statistical offices have created special units aimed at servicing special user categories. In view of the technological developments, it may be a more feasible solution to organise a user-oriented clearing-house (as a single unit) with a flexible and open-ended infrastructure.

The second chapter provides an extensive analysis of the following major types of SIS:

- survey processing systems,
- clearing-house systems, "data warehouses",
- registers,
- analytical processing systems.

The author reviews the tasks, functions and requirements of each of these major information systems. Special attention is drawn to the survey processing systems covering the full life-cycle of a statistical survey: its planning, operation and evaluation.

During the planning phase, the designers of the survey make decisions concerning the major purposes and users of the survey, major inputs and outputs, procedures for obtaining the inputs and transforming them into outputs. It is useful if the designers of a statistical survey have access to a knowledge base, containing information about the design of similar or related surveys. To enable to learn from the experiences gained, all-important information on statistical survey design should be documented. Metadata on quality and contents

of data and processes, and feedback from users are a very important part of such documentation.

The survey operation phase consists of the following main processes: frame creation, sampling, measurement, data preparation (data entry, coding and data editing), creation of the observation register, estimation, analysis and the presentation and dissemination of results. The estimation process is often combined with production-oriented analysis aimed at improving the quality and efficiency of future surveys.

The results of the survey should be made available in a user-acceptable form via appropriate distribution channels. In principle, a survey production cycle is completed once the results of the survey have been published. A trend to include the electronic dissemination of the results in the publishing concept can be observed. Important components of the new publishing system could be the clearing-house function and Internet.

The survey evaluation consists of checking and evaluating whether the specified end-products have been delivered, the outputs properly published and advertised, the metadata documented and stored, and of the assessment of the production-oriented metadata and user feedback.

The register function lies in maintaining up-to-date information on all objects belonging to a certain population. The registers can serve as sampling frames for surveys. In addition to maintaining the current status of the population, the register should permit the reconstruction of the population of objects at any point in time, and to reproduce the original status and all events that have affected the objects. A special kind of registers are those containing metadata, such as definitions, links to surveys and data sets, standard formats, value sets, etc.

The clearing-house function facilitates the exchange of data and metadata between different surveys, registers and analysis functions, including external users. Another label for this function is "data warehouse". The clearing-house function receives and delivers

data and metadata according to specified standard formats, following specified delivery procedures.

In addition to the production-oriented analysis mentioned above, statistical offices perform some more user-oriented analysis. When such an analysis uses data from several statistical surveys and other sources, the analytical processing system can be regarded as a separate system with interfaces to the survey production systems.

The third chapter outlines a future information systems architecture for a statistical organisation. It is based upon the four major types of statistical information systems specified in the previous chapter. An important component of the proposed architecture is a corporate data warehouse, encompassing all clearing-house functions and register functions.

The future corporate data warehouse of a statistical organisation includes five compartments:

- raw data and metadata;
- final observation registers;
- final multidimensional statistics;
- electronic documents;
- global metadata, including registers.

Data and metadata in the raw data compartment will not always be in a standardised form. There should be generalised software supporting the standardisation of data and metadata. In addition, there should be generalised software tools supporting all important processes and sub-processes in survey processing systems and analytical processing systems.

In the case of international organisations, most of the member countries deliver data and at least some metadata electronically. Even so, the data may arrive in many different formats. Thus, a first step will be to standardise incoming data and metadata. This step can be avoided, if member countries agree to provide data and metadata according to some international standard, e.g. the EDIFACT standard for

GEneric Statistical MESsages (GESMES). It is important to note that a standard format must include standards for both data and metadata.

Chapter 4 analyses the technical aspects of the proposed architecture. As the statistical information systems should provide information for many different kinds of users, with different and sometimes contradictory needs, flexibility is a particularly important consideration for all the hardware, software and data components.

A statistical office very often runs a relatively large number of different statistical applications. However, many of these applications are rather similar in the sense that they perform a limited number of functions, which are typical for information systems supporting statistical surveys, i.e. survey planning, survey operation and survey evaluation functions. When a statistical function or subfunction is analysed, at some stage a level is reached at which the software components need not necessarily be tailored to the needs of statistical applications. Instead, general-purpose standard software components may be used. Nowadays, this "general purpose level" may appear rather high up in the systems architecture of a statistical application.

The same principle of preferring standard and re-usable components applies to the hardware. The IBM compatible PC has long since become a de facto standard hardware component for statistical organisations and for the users and customers of statistical organisations.

The data components of an information system are stored either as physically integrated parts of the application software system or as separate files or databases. Program/data independence is an important requirement meaning that the software and data components of an information system may be developed and maintained relatively independently of each other. A modification of the contents, structure, or storage of data should not necessitate modifications of programs using the data. On the other hand, it should be possible to modify or add software components without having to redefine data components.

Analysis of different information systems architectures leads to a proposal for a multi-tier network-based information systems architecture that balances the needs for centralisation and decentralisation in a modern statistical organisation.

Chapter 5 focuses on the implementation aspects of the proposed IT architecture. Under the conditions of rapid IT development, there must be a realistic plan for implementation which is able to accommodate changes that happen during the implementation process itself. Some recommended development principles are the following:

(i) as the price/performance ratios of standard hardware and software improve all the time, it is better to buy standard components off the shelf rather than develop one's own solutions, and to spend more on hardware capacity rather than complicating a simple software solution;

(ii) it is safer to standardise in terms of interfaces between components rather than in components themselves;

(iii) instead of waiting for better standards, better hardware and software, buy the state-of-the-art hardware and software components, and replace a component with a better one as soon as possible, without having to change any other components;

(iv) have a clear picture of an organisation's overall information systems architecture and define a number of strategically important interfaces;

(v) the maximum time-frame for development projects is not more than two years, complex projects should be divided into subprojects with clearly defined results and deadlines, too many and over-ambitious goals should be avoided;

(vi) while migrating to a new technical platform, one can take the opportunity to improve the contents and quality of statistics at the same time but only to the extent that such activities do not threaten the time schedule of the project; depending on how important these improvements really are, some deficiencies might be acceptable and enhanced in the long run by sustainable improvement.

The role of top management in this process is essential. However, top management needs support from the subject-matter statisticians. The project should focus on statistical tasks; IT serves as a major instrument which the project has at its disposal. Possibilities to improve statistical co-ordination should be noticed and actively exploited, e.g. by means of the global metadata component of the data warehouse.

CHAPTER 1

STATISTICAL ORGANISATIONS

A statistical organisation is an organisation where production and analysis of statistics are important parts of the work. National statistical offices as well as international organisations such as Eurostat and the OECD (Statistics Directorate) are typical examples of statistical organisations in this sense.

1.1   Statistical information systems

A statistical information system performs certain typical processes, statistical processes, and it handles certain typical categories of data, statistical data. Furthermore, a statistical information system is associated with certain typical categories of users and purposes.

1.1.1   Statistical processes

In a statistical organisation there are processes for the following types of tasks:

- collecting statistical data (microdata, macrodata, metadata),
- processing statistical data,
- storing statistical data,
- retrieving statistical data,
- analysing statistical data,
- disseminating statistical data.

Such processes are called statistical processes. Statistical processes use and produce statistical data. Statistical data may be microdata, macrodata and/or metadata.

1.1.2   Statistical data: microdata, macrodata and metadata

Microdata are data about individual objects (persons, companies, events, transactions, etc). Objects have properties which are often expressed as values of variables of the objects. For example, a "person" object may have values of variables such as "name", "address", "age", "income". Microdata represent observed or

derived values of certain variables for certain objects.

Macrodata, "statistics", are estimated values of statistical characteristics concerning sets of objects, "populations". A statistical characteristic is a measure that summarises the values of a certain variable of the objects in a population. "The average age of persons living in OECD countries" is an example of a statistical characteristic. Some statistical characteristics, e.g. correlations, summarise the values of more than one variable. Macrodata represent estimated values of statistical characteristics. Estimated values deviate from true values because of different imperfections (errors and uncertainties) in the underlying observation (measurement) and derivation processes. The difference between "estimated" and "true" values is an issue not only on the macro level, but also on the micro level, since the observed (measured) values deviate from the true values because of measurement errors.

Statistical metadata are data describing different quality aspects of statistical data, e.g.

- contents aspects, describing definitions of objects, populations, variables, etc;

- accuracy aspects, describing different kinds of deviations between observed/estimated and true values of variables and statistical characteristics;

- availability aspects, describing which statistical data are available, where they are located, and how they can be accessed.

1.1.3   National and international statistical organisations

In a particular statistical organisation there may be an emphasis on certain types of statistical processes, whereas other types of processes are less important, or even non-

existing. For example, in an international statistical organisation such as Eurostat or the OECD, there is a natural emphasis on analysing and comparing statistics and metadata produced by statistical organisations in member countries of the organisation. International organisations do not usually collect their own microdata.

On the other hand, the concepts of "macrodata" and "microdata" are relative concepts in a certain sense. For example, macrodata about populations of persons and companies, delivered by national statistical offices, may rightfully be regarded as microdata about individual countries by an international organisation. If this latter view is taken, statistical organisations in member countries will become respondents to surveys conducted by an international organisation, and the member countries themselves will become observation objects, in the same way as persons and companies are observation objects in surveys conducted by a national statistical office.

1.1.4 Applications and infrastructures

An information system, where most processes are statistical processes in the sense stated above, is called a statistical information system (SIS).

Typically, an information system is associated with a specific purpose, e.g. to produce a specific information product, like a certain data set and/or a certain report or publication. Such an information system is called an information system application, or an application, for short.

A typical example of an application in a national statistical office would be the information system associated with a statistical survey conducted regularly, for example, the Labour Force Survey (LFS). Internationally, an analogous example would be the OECD information system for producing the Main Economic Indicators (MEI).

In addition to information system applications with very specific end products, there are information systems with more general

purposes, e.g. the information system associated with a database service to the external and/or internal users of the statistical data produced by a statistical organisation. As a matter of fact, it is a typical feature of statistical information in general, and official statistics in particular, that the information is "multi-purpose" and that many of the specific users and usages are unknown at the time when the statistics are designed and produced.

1.1.5 Typical flow of data and metadata through a statistical survey

An information system (or a system of information systems) that provides a general service to end-users, or to other information systems, is sometimes called an information system infrastructure. A well-designed information system infrastructure could also serve as an efficient basis for further application developments. If all the information systems of an organisation are well designed and co-ordinated within some kind of common framework for the organisation, they form together *the* information system infrastructure, or *the* information system, of the organisation.

1.2 Information system architecture of a statistical organisation

The information system architecture of an organisation is a common framework, within which different kinds of individual information systems play their respective roles and interact with one another.

1.2.1 Some general purposes of an information system architecture

The information system architecture of an organisation is a framework for structuring and co-ordinating

- the subsystems and components of the individual information system applications;

- the interaction between different information system applications;

- the subsystems and components of the information system infrastructure;

NATIONAL STATISTICAL OFFICE

INTERNATIONAL ORGANIZATION

A typical frame is a register of companies (= observation objects) and contact persons (= respondents)

A typical frame is a register of national statistical offices (and other data providers) with contact persons.

Electronic or paper questionnaire or other measurement instrument

Questionnaire or format for electronic delivery of data and metadata.

Instructions to respondents and/or interviewers

Instructions to respondents.

Frame

Questionnaire

Instructions

Measurement and registration

Respondents complete questionnaires and send them by mail, voice or electronic communication to the statistical office.

Respondents in member countries deliver statistical data and metadata to the international organization.

Raw data

Metadata

Examples of metadata accompanying raw data are: questions asked, instructions, comments by respondents and interviewers, answers to metadata questions.

Statistics from member countries.

Metadata concerning statistics from member countries.

Data preparation

Data in progress

Coding and editing of raw data and metadata, based upon coding and editing rules. Register and data from other sources, including data from earlier repetitions of the same survey, may be used.

Checking input data and metadata for possible errors and incompletenesses. Adding data and metadata from other sources.

Final observation register

The final observation register of a survey contains "clean" microdata and accompanying metadata, organized in a standardized way in order to facilitate subsequent processes.

The final observation register of a statistical product contains "clean" statistical data and metadata from member countries, organized in a standardized way.

Estimation and analysis

Data in progress

Computing estimated values of statistical characteristics for populations and subpopulations. Performing statistical analyses.

Computing derived data, updating time series data, peforming seasonal adjustments and other analyses.

Final statistics

Final multidimensional statistics and accompanying metadata, organized in a standardized way.

Final multidimensional statistics (cross-sectional data and time series data) with accompanying metadata, organized in a standardized way.

Presentation and dissemination

Final statistical products are compiled from data and metadata in the final statistics database. The products may be stored as electronic documents, and they are disseminated via different channels and media.

Final statistical products are compiled from data and metadata in the final statistics database. The products ma be stored as electronic documents, and they are disseminated via different channels and media.

Printed publications

Electronic documents

CDs and diskettes

Figure 1.1   Typical flow of data and metadata through the processes of a survey of a national statistical office (explanations on the left side of the flow) or a statistical product of an international organisation (explanations on the right side of the flow).

- the interaction between the applications and the infrastructure.

For example, information system architecture may define

- a standard functional structure for the subsystems and components of a certain application type, e.g. a statistical survey application;

- standard interfaces for interactions and data exchange between different types of sub-systems or components;

- standard interfaces for human interactions with information systems;

- standard interfaces for interactions and data exchange with external information systems;

- services to be provided by the information system infrastructure;

- standard hardware, software, and data components to be used.

1.2.2 Information system architecture vs organisation architecture

Ideally, the information system architecture of an organisation should be in harmony with the architecture of the organisation as such, and the architecture of the organisation should be a reflection of the purposes and tasks of the organisation.

The way the organisation has chosen to organise itself may have important implications for the design of its information system architecture. For example, if the organisation has chosen a highly decentralised control system for a certain type of work, it would not be appropriate for the information systems supporting this work to require highly centralised control. On the other hand, even in a decentralised organisation, there are certain communication and co-ordination needs that should be supported by a common information

system infrastructure. However, as is exemplified by the Internet, such an infrastructure in itself need not necessarily require a very centralised control. The important thing is that there are certain widely accepted *de facto* standards as regards communication interfaces and other important matters.

Consequently, in order to analyse and (re) design the information system architecture of a statistical organisation, we must pay attention to the purposes and tasks of the statistical organisation, as well as to the way the statistical organisation has chosen to organise its functions and processes.

The traditional building blocks of a national statistical office are the statistical surveys performed by the organisation. Each survey is associated with (at least) one data collection, and it produces a certain subset of the official statistics of the particular country, usually according to a repetitive schedule (monthly, quarterly, yearly).

Analogously, the statistical activities of an international statistical organisation are organised around its "surveys" or product-oriented applications. The data inputs of these "surveys" on the international level are statistical outputs from national statistical offices in member countries, and the outputs of the international surveys are the official statistics of the particular international organisation.

Thus statistical organisations are traditionally organised by surveys. Since surveys are typically defined by data collections, this means that statistical organisations are "input-oriented" in a certain sense. A statistical survey is often based upon one major data collection process (nowadays often supplemented with input data from other sources, including administrative registers), and the survey is typically handled "from grain of wheat to baked bread", from beginning to end, by one and the same organisational unit. Sometimes certain functions, or subfunctions, like data entry, systems development and

programming, computer operations, and printing and publishing, are performed by separate functions within or even outside the statistical organisation, but even so, it is usually the individual surveys that have the overall responsibility for the end-products of the statistical organisation. The survey-defined organisational units are in control, and even top management sometimes has difficulties achieving desirable co-ordination and control across surveys and subject-matter areas.

And there are indeed very strong and legitimate needs for cross-survey co-ordination and control. Users of statistics typically have information needs across data collections and across surveys, i.e. across the organisational boundaries within the statistical organisation, which are based upon surveys and subject-matter areas. When we talk about "users" here, we are primarily thinking of "external users", that is, users outside the statistical agency. In addition to external users there are internal users, and they may have different needs. Thus there is a requirement for clearing-house functions within a statistical organisation, matching the requirements of the customers with the possibilities provided by the surveys.

Statistical organisations which have become aware of the drawbacks of their traditional input orientation have sometimes chosen to add organisational units oriented towards special user categories, such as regional planners, researchers, schools, etc. A problem here is that users are not so easy to classify into distinct categories, as are surveys and data collections. Whichever classification of users is chosen, there will always be a lot of overlaps and unresolved conflicts regarding user needs. Particularly in view of the technological developments that have taken place over the last few years, it may be a more feasible solution to organise only one major user-oriented clearing-house function. By necessity, such a general clearing-house function must have the character of a very flexible and open-ended infrastructure, which can be used directly by many users themselves (e.g. self-service via the Internet). Naturally, such a user-oriented infrastructure must also involve numerous experts from

surveys and subject-matter areas, who can assist users with more complex problems.

Whenever a statistical organisation is exposed to cross-survey usage of statistical data, it will also become more aware of the needs of certain input-oriented co-ordination activities. In order to enable the users to combine data from different surveys in a responsible way, the statistical organisation must co-ordinate definitions of observation objects and observation variables. Registers and classifications are the traditional tools for improving such co-ordination. It is not unlikely that these tools, and the organisational units responsible for these tools (not only from a technical point of view), will play a more active and visible role in the future. We shall use the term "register functions" as a common label for those functions responsible for registers, classifications, and other content-oriented (and technical) co-ordination tools, such as catalogues of variables, or so-called data dictionaries. It should be stressed that a register function within a statistical agency does not have the same role as a register function in an administrative agency. In an administrative agency the purpose of a register is to maintain updated information about objects of a certain kind for the purposes that the administrative agency is responsible for, for example collecting taxes or paying social benefits. In a statistical agency the purpose of a register is to support statistical tasks, such as establishing frames for surveys, and co-ordinating surveys. Administrative registers can be used as sources for statistical registers.

Thus we have identified three major types of functions in a statistical organisation:

- survey functions,
- clearing-house functions,
- register functions.

In many statistical organisations there is a fourth type of fundamental function:

- analysis functions.

In some international organisations, like the OECD, this type of function is the most fundamental one. In national environments it is common for important analysis functions to become (statistical) organisations in their own right, separate from the national statistical offices.

In summary, the statistical information system architecture of a statistical organisation should provide an efficient common framework for individual information systems corresponding to

- survey functions,
- clearing-house functions,
- register functions,
- analysis functions.

Furthermore, the common framework should promote efficient co-ordination and data and metadata exchange within and between different statistical information systems. The statistical information systems architecture should make individual statistical information systems co-operate in such a way that together they appear as one integrated, and yet open-ended, system, one flexible, efficient, and powerful information infrastructure, to the benefit of internal and external users.

CHAPTER 2

FOUR MAJOR TYPES OF STATISTICAL INFORMATION SYSTEMS

We shall now analyse the tasks, functions, and requirements of each one of the four major types of statistical information systems identified in the previous section:

- survey processing systems,

- clearing-house systems, "data warehouses",

- registers,

- analytical processing systems.

2.1   Survey processing systems

The life cycle of a statistical survey consists of three major phases:

- survey planning,

- survey operation,

- survey evaluation.

If (more or less) "the same" survey is repeated regularly, e.g. monthly, quarterly, or yearly, the life-cycle is repetitive, too. When the survey is carried out for the first time, many important aspects have to be carefully planned "from scratch" (although some valuable lessons may have been learnt from the running of other surveys). After this first-time planning phase, the survey may often be repeated many times without too much (re)planning before every (re)iteration of it. However, after each iteration of the survey, there should be an evaluation of experiences gained, and these experiences should be recognised as important for the planning of future iterations of the survey, as well as for the interpretation of the data emanating from the survey iteration that has just taken place.

2.1.1   Survey planning

During the planning phase, the designers of a survey make at least preliminary decisions concerning

- major purposes and users of the survey;

- major outputs from the survey, needed to satisfy the major purposes and user needs;

- major inputs to the survey that are needed for the production of the required outputs;

- main procedures for obtaining the inputs and transforming them into the outputs.

The preliminary decisions may become revised during the subsequent operation and evaluation phases.

It is useful for the designers of a statistical survey if they have access to a knowledge base containing information about the design of similar or related surveys. Traditionally, this type of knowledge is mostly in the heads of the designers themselves, or in the heads of those colleagues of the designers who happen to be around. However, today there are possibilities to develop this type of design support in a more systematic way, using computerised metainformation systems and expert systems.

In order to actively support a self-learning organisation, the management of a statistical organisation should insist that design decisions concerning statistical surveys, as well as experiences from executing the designs, be systematically documented and made available for future users and (re)designers. Metadata concerning quality and contents of data and

processes are important parts of such a feedback system, as are user evaluations.

Major purposes and usage of a statistical survey

The purposes and usage's of a statistical survey can be divided into primary and secondary purposes/usages.

Primary purposes and usage's of a statistical survey

A typical primary purpose of a statistical survey is

- to provide estimates of known quality of certain statistical characteristics, needed by certain users for better understanding a certain phenomenon, and for solving problems related to this phenomenon.

Secondary purposes and usage's of a statistical survey

In addition to the primary purpose of a statistical survey, which is often relatively well defined, it is usually assumed that the statistical survey should also satisfy a number of secondary purposes. These may not be so easy to specify since they often refer to future needs, which are not yet known at all, or at least not known in very great detail.

A typical secondary purpose of a statistical survey is to provide statistical data (microdata, macrodata, and metadata) that can be used by various users, now or in the future, for providing

- estimates of known quality of *other statistical characteristics* than those specified by the primary purpose of the survey; and/or

- *differently defined estimates* of known quality of the same statistical characteristics as those specified by the primary purpose of the survey.

In order to fulfil the criterion that the estimates should be of known quality, they must

be accompanied by a certain amount of metadata. Even more metadata are needed in order to ensure that future users can actually (re)use stored statistical data for new purposes. We will return to these problems.

Major outputs from a statistical survey

There are three major categories of outputs from a statistical survey:

- *macrodata*, "statistics", representing estimates of known quality of certain statistical characteristics; these data are essential for the primary purpose(s) of the survey;

- microdata, "observations of individual objects", underlying the macrodata produced by the survey; these data are essential for (future) (re)users and (re)interpreters of the survey results;

- *metadata*, "data describing the meaning, accuracy, availability, and other important aspects of the quality of other data"; these data are essential for correctly identifying and retrieving potentially relevant statistical data for a certain problem, as well as for correctly interpreting and (re)using statistical data.

Traditionally, macrodata are published and stored in statistical tables in printed publications. No doubt these traditional forms and channels for presenting survey results to users will continue to exist for the foreseeable future, but more and more users will prefer, and strongly demand, electronic alternatives. There are emerging standards for storing statistical data in certain types of multi-dimensional structures, sometimes called (multi-dimensional) boxes or cubicles. Here we shall call them multi-dimensional tables.

Although the traditional forms and channels for disseminating statistics will continue to exist in the future, the production of such outputs will typically be based upon electronic outputs rather than the other way around. For example, a statistical output like the Main Economic Indicators of the OECD will first be stored and

published through a database service, and only then printed versions of the same results will be produced, and in addition the printed results will be produced automatically from the data and metadata stored in the database.

Microdata from statistical surveys will be stored and made available (under confidentiality restrictions) in standardised electronic form, e.g. as relational tables in relational databases, or as so-called "flat files". The versions of the survey microdata that will be stored for future (re)use are the so-called final observation registers, containing observation data that have been edited and possibly "enriched" with data from related surveys. Final observation registers should never be updated. If a need should occur to "revise" the data of a final observation register, a new final observation register should be created, and the old one should continue to exist.

It should be remembered that for an international organisation, "microdata" will usually mean "data reported by statistical organisations in member countries". In the member countries these data are, of course, themselves macrodata that have been aggregated from (another) micro-level of observation data, such as data about individual companies. (It could be noted that these observation data again are probably, from the perspective of the individual company, the result of another aggregation process within the company.)

Metadata from a statistical survey should be stored in such a way that, from a user's point of view, they are fully integrated with the data that they describe. In addition to these "survey-local" metadata, there is a need for maintaining global metadata holdings. The global metadata holdings will typically be principal components of the three other types of statistical information systems (clearing-houses, registers, and analytical systems).

Major inputs to a statistical survey

Traditionally, a statistical survey is associated with one major data collection, and this data collection is typically obtained by means of some kind of questionnaire with questions (metadata + slot for data) and instructions (metadata), which is either sent to the respondent for self-administration, or used by special interviewers, visiting or phoning the respondents to obtain answers to the questions, i.e. data for the empty data slots in the questionnaire.

Nowadays, a statistical survey may often obtain input data from several sources. A register is often used as a frame (sometimes a sampling frame) for identifying and locating the respondents and/or the observation objects, and this register may itself contain data which are also useful for the statistical purposes of the survey. In some countries, with well-developed administrative information systems concerning taxes and other matters, the data obtained directly by the survey as such may be enriched with data from relevant administrative sources. The survey data may also be enriched with data from other statistical surveys, and with data that are derived from other data, according to formal definitions, represented by computerised algorithms.

Main procedures for obtaining inputs and transforming them into outputs

In the next section (section 2.1.2 below) we shall analyse the typical operational processes in a statistical survey. During the planning phase, the designers take at least preliminary decisions, determining "the gross picture" of these procedures.

The most characteristic procedure of a statistical survey is the procedure whereby observation data are transformed into estimated values of statistical characteristics, i.e. the procedure whereby microdata are aggregated into macrodata. This procedure is called estimation. In order to understand the essence of statistical estimation we need a precise definition of a statistical characteristic.

What is a statistical characteristic?

A statistical characteristic is defined by a triple

- $<O, V, f>$

where

- O is a set of objects (or object vectors), called a population;

- V is a variable (or a vector of variables) having values for the objects in the population;

- f is an operator, called a statistical measure, producing a value f(O, V) for the population from the values of the variables for the objects in the population.

Typical examples of statistical measures are frequency count, sum, average, and variance.

The population is often structured into subpopulations, for which estimates are produced as well.

Time usually plays an important role in the definition of a statistical characteristic. The population is often defined as the set of objects of a certain type, having a certain property (or combination of properties) in common at a certain point of time. Alternatively, the population can be defined as the set of objects of a certain type that have been born, lived, or died during a certain time period, e.g. the events of a certain type that have occurred during a certain year, or the processes of a certain kind that have started, been on-going, or stopped during a certain month.

The variable V must usually be qualified by a time parameter, too, in order to ensure that every object in the population is associated with a unique value (or set of values, in the case of multi-valued variables). If V is a set of variables, all the variables may be separately qualified by (possibly different) time parameters.

Some examples of statistical characteristics:

- the number of people living in Canada at the end of 1996;

- the average income of people living in France at the end of 1996;

- the total value in current US dollars of the production of commodities in the United States during the first quarter of 1996;

- the number of road accidents that have occurred in Germany during 1996;

- the average length of hospital treatments that was on-going in Holland during (at least) some part of 1992;

- the average percentage increase/decrease between 1995 and 1996 in the annual income of people living in Sweden during the whole of the two-year period 1995-1996.

The (true) value of a statistical characteristic is derived (by means of an aggregation process) from the (true) values of one or more sets of object characteristics. The estimated value of a statistical characteristic is derived (by means of another aggregation process, called the estimation procedure) from the observed values of (possibly the same) sets of object characteristics, the so-called observation characteristics.

What is an object characteristic?

An object characteristic is defined by an ordered pair

- $<O, V>$

where

- O is a set of objects (or object vectors), called a population;

- V is a variable (or an object relation), having values for the objects in the population.

Time plays a similar role in the definition of an object characteristic as it does for the definition of a statistical characteristic.

Each object (or object vector) in the population is associated with one instance of the object

characteristic. At any particular time t, each object (or object vector) in the population is associated with a unique value of V (or with a unique set of values of V in the case of multi-valued variables).

Summary of the survey planning phase

In summary the survey-planning phase consists of the following main processes:

- Specifying the survey contents in terms of (i) the statistical characteristics to be aimed at by the estimation procedures of the survey, the so-called target characteristics; and (ii) the observation characteristics to be observed, directly or indirectly, by the data collection procedures of the survey;

- Establishing the overall design of the survey procedures in terms of (i) data collection procedures (including frame creation, sampling, measurement, data preparation, and observation register creation); (ii) procedures for estimation and certain types of analyses; and (iii) procedures for presentation and dissemination.

2.1.2 Survey operation

The survey operation phase consists of the following main processes:

1. **Frame creation:** Establishing a frame in the form of a register (see section 2.2). The frame is used for identifying and locating sources/respondents and/or observation objects.

2. **Sampling:** In the case of a sample survey, a sample of sources/respondents and/or observation objects is drawn from the frame according to certain rules, the sampling procedure.

3. **Measurement:** The observation objects are observed/measured with respect to a number of observation variables. This results in observation data, so-called primary data or raw data. Observation data are collected or received, directly or indirectly, from respondents. Measurement procedures typically make use of measurement instruments, like questionnaires. Questionnaires (paper-based or electronic) may be self-administered by the respondents or handled by interviewers visiting the respondents, or contacting them by phone.

4. **Data preparation:** The raw data are prepared for statistical estimation by means of certain typical processes like data entry, data coding, and data editing. Rules are often used for formalising and controlling the procedures used. The data preparation processes are often combined with each other. Sometimes at least parts of the data preparation processes are integrated with the measurement process, where the raw data are collected from the respondents; for example, this is typically the case if an interviewer is equipped with a computer containing the questionnaire and some rules and background data that can be used for immediate coding and editing of the data captured from respondents.

5. **Observation register creation:** The incoming observation data are stored and organised in an observation register for the particular (repetition of) the survey. The observation register is typically organised in a certain standard way, e.g. as a relational database, or as a set of flat files, so that standard commercial software packages can easily be applied on the observation data. Before the raw data have been collected, the observation register may be an empty structure, or it may contain some data copied or derived from the survey frame and/or other sources that already exist. As a result of the data collection process, the observation register is step-by-step filled with data. Coding and editing operations may cause the observation register to be updated. Finally, the observation register is closed and delivered to an archive. By definition, this final observation register will never be updated again. If someone wants to change the contents of a final observation register, he or she will first have to make a copy of the final observation register, thus creating another observation register.

6. **Estimation and analysis:** Estimated values of certain statistical characteristics, the target characteristics of the statistical survey, are

derived from the observations of observation characteristics that are stored in the observation register. The estimations are based upon estimation rules, forming estimation procedures. An estimation procedure is dependent on the sampling rules as well as on various assumptions about populations, measurements and non-response. The estimation process is often combined with certain types of analyses. Most analyses of survey outputs are carried out by users outside the survey organisation. However, some user-oriented analyses and a lot of production-oriented analyses are carried out in direct connection with the survey. A major purpose of production-oriented analysis is to improve the quality and efficiency of future surveys.

7. **Presentation and dissemination:** The results from the survey (final observation registers, with metadata, as well as aggregated statistics, with metadata) are made available in suitable forms, e.g. statistical tables and graphs, and on suitable media, e.g. paper publications and electronic databases, and are disseminated through suitable channels, e.g. through publishers, fax, and the Internet.

A note on frames in an international environment

Frames that are directly used by international statistical organisations are relatively simple and uncomplicated, consisting mainly of lists of member countries and organisations and contact persons within member countries that provide input data and metadata to the international organisation. Nevertheless, these inputs from the member countries may themselves be dependent on much more complicated frames and frame-related procedures. In order to judge the quality of data from member countries, in particular as regards coverage, the statistical staff of the international organisation needs to have a certain knowledge about frames and frame-related problems.

Basically, a survey frame is a list of objects, which is used for

- sampling, if the survey is a sample survey;

- identifying and locating the respondents, i.e. the objects from which data are going to be collected or received.

If the survey is a total enumeration, a so-called census, the set of objects listed by the survey frame may be identical with

- the set of respondents;

- the observation population of the survey, i.e. the set of objects to be observed;

- the target population of the survey, i.e. the population of objects for which certain statistical characteristics are going to be estimated by means of the data collected by the survey.

However, in more complicated surveys the set of objects listed by the survey frame may be different from each one of these three sets of objects, which in turn may be different from each other as well. However, there must always be well-defined relationships between all these sets of objects, so that

- the frame can be used for identifying and locating the respondents;

- data about the observation objects can be obtained or derived from the respondents;

- estimated values of statistical characteristics of the target population (and subsets of the target populations) can be derived from the observed or derived data about the observation objects.

Similarly, in a sample survey, a sample of objects is obtained from the frame by means of a sampling procedure, and the sample must be related in a well-defined way to the respondents, the observation objects, and the target objects.

A note on observation registers

At any stage between the raw data stage and the final observation register stage, an observation register may be temporarily

"frozen" and used as the basis for statistical estimation and analysis. The purpose of these statistical processes may be production-oriented and/or user-oriented. Production-oriented estimation and analysis particularly may lead to the detection of probable errors in the observation register, and this may lead to an update of the observation register, which will then be in a new stage. Whenever estimation and analysis leads to publishing and/or permanent storing of statistics, the underlying observation register should be stored as well, so that a subsequent user of the statistics may question and change the assumptions made by the statistician/analyst who provided the original estimates and analyses - in accordance with sound scientific principles. If it is the final observation register that is used for estimation and analysis, this condition will be automatically fulfilled, since the final observation register is archived and will not be changed, by definition.

Within an international statistical organisation there may be relatively frequent needs for temporary freezing of observation registers, since different member countries may deliver their data at different dates, and some countries may be lagging behind. Further processing and analysis should not be delayed until all countries have responded. This is especially the case when the work is organised by country.

A note on publishing, presentation, dissemination, and analysis

In principle, a survey production cycle is completed once the direct results of the survey have been published. Publishing may be defined in different ways. Traditionally, statistics were published by means of printed publications, reports or newsletters, containing aggregated statistical data together with some metadata and simple analyses. Now there is a trend towards redefining the publishing concept by associating publishing with making the direct results from the survey electronically available to the customers/users (in the case of official statistics: the public at large), e.g. through the clearing-house function and/or via the Internet

(which may be part of the clearing-house function).

The term "analysis" is difficult to define. We shall use it as a practical label for everything that comes after (first-time) publishing of statistical data. As was just indicated, (first time) publishing itself may be associated with a certain amount of (usually rather simple) analyses.

Statistical analysis will often involve statistical data (macrodata, microdata and metadata) from several types and repetitions of surveys performed by the statistical organisation, as well as statistical and non-statistical data from other sources. Many analyses of the statistical data from a statistical organisation are carried out outside the statistical organisation, by many different types of users of statistical data: research institutes, participants in political processes, financial institutions, massmedia, schools, etc.

The final results from statistical analyses that take place inside the statistical organisation itself should typically be published, stored, and disseminated through the clearing-house function. Some production-oriented analyses may not be published, since the purpose of such analyses is to improve the quality of the survey production processes. Nevertheless, the results of such analyses, too, should be recycled through the clearing-house function; they will constitute a valuable part of the metadata associated with the statistical data produced by the statistical organisation and made available to users through the data warehouse.

2.1.3   *Survey evaluation*

During the evaluation phase certain things should be checked and evaluated. For example, the following should be verified:

- specified macrodata end-products have been delivered, accompanied by specified metadata, to the organisation's data warehouse function, if such a function exists;

- specified microdata end-products, final observation register, have been delivered, accompanied by specified metadata, to the data warehouse function, if such a function exists;

- specified outputs (including database outputs as well as printed material) from the survey have been properly published and advertised;

- production-oriented metadata have been properly documented and stored, so that the quality and efficiency of the survey outputs and the survey processes can be kept under control, now and in the future;

- production-oriented metadata, as well as feed-back data from the users, have been properly evaluated, so that appropriate replanning of survey contents and procedures can take place before the next operation phase of the survey, if the survey is going to be repeated.
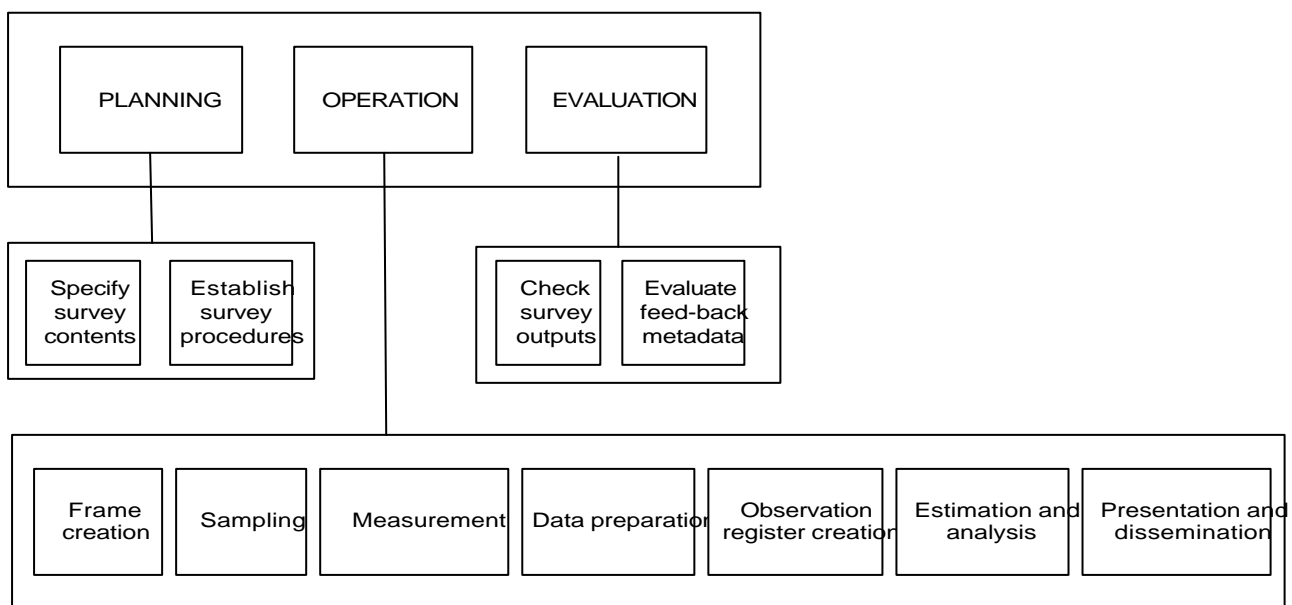
2.1.4   Summary of survey processing systems

Figure 2.1 visualises the life cycle and the processes of a survey processing system.

2.2   Registers

The basic purpose of a register is to provide an authorised, correct and up-to-date listing of all entities of a certain kind. The entities may be real-world objects such as persons, companies, cars, countries, or may be meta-objects such as variables and values. The term "register" is often reserved for object registers, i.e. registers of real-world objects. In such cases, registers of variables, values and other meta-objects may have labels like "catalogue of variables", "classification database" and "data dictionary".

2.2.1   Object registers

In statistics production (object) registers are used for maintaining up-to-date listings of all objects belonging to a certain population, for example "all persons living in Sweden". Such registers are used as survey frames. If the survey is a sample survey, a register will typically be used as a sampling frame. In addition to object identifiers, this kind of register typically contains information for locating the objects, as well as other basic information about the objects which can be used for stratification and sampling, and even as a basis for producing estimates of certain basic statistical characteristics, i.e. for simple aggregations and tabulations.

Figure 2.1 A visual summary of the phases and processes of a survey processing system

An international statistical organisation will need an object register containing

- names and codes for member countries, other countries and combinations of countries and parts of countries for which the organisation collects data;

- names, addresses, phone and fax numbers, e-mail identities and so on for respondents (organisations and persons) to the surveys conducted by the international organisation;

- other basic facts about respondents and observation objects of surveys conducted by the international organisation.

Generally speaking, an object register function should

- maintain a current version of the register, which is as correct and up-to-date as possible;

- provide "time t" versions of the register based upon what is known at "time t+d" where t are certain prescribed time point (like the end of each year or month), and d is a certain prescribed delay;

- be able to reconstruct the population of objects at any point of time t in the past as correctly as possible, taking into account all events that have happened up to t, according to transactions received up to current time (or any time point between t and current time);

- be able to reproduce the original status and all events that have affected the objects in the register population, as reported by the transactions that have been received by the register function.

In the case of an international organisation, an object register function needs to take care of events such as

- new member countries entering the organisation,

- member countries leaving the organisation,

- territorial or other changes in the definition of a member country or some other object of interest (e.g. a combination of countries like the EU or NAFTA).

In connection with such events it may be necessary to reconstruct old statistics to make them comparable with forthcoming data. Such reconstruction may require special procedures and tools, and it would be appropriate for a register function to provide such tools, whenever needed.

2.2.2   Meta-object registers

A variable register (or a catalogue of variables, or "data dictionary", as it is often called) contains a listing of the variables used by an organisation. In addition to variable identifiers, this kind of register typically contains information about the definition of the variable, links to surveys and data sets, where the respective variables occur, and standard representation formats for values of the variable, when stored on electronic media.

One way of defining a variable is by referring to the question asked (including instructions to the respondent and/or interviewer) when obtaining the value of the variable for an observation object. Another important part of the definition of a variable is the set of valid values for the variable, i.e. the set of valid answers to the question asked, before and/or after coding the answers. The set of valid values of a certain variable is called the value set of the variable.

Different variables may share the same value set. For example, "birth place" and "place of residence" are two different variables for persons that may share a value set-containing identifiers of "places" (e.g. countries, counties, communes). Similarly, "the age of a person" and "the age of a person's mother" are two different variables, which may share the same value set of valid "ages".

Actually, a value set is in itself a register, since it can be regarded as an authorised, up-to-

date listing of the values that can be taken by certain variables. In addition to value identifiers (value codes), this kind of register typically contains names (text labels) for the values that are suitable for presentations, e.g. in the stubs and headings of statistical tables.

The value sets of all variables used by an organisation may be combined into a superset of values; any particular value set of any particular variable will be a subset of such a super value set, which is again a register in the sense that it contains an authorised, up-to-date listing of the values that can be taken by variables used by the organisation. In addition to value identifiers and text labels, this kind of register would contain links to value sets and variables.

A register is often associated with, or even combined with, one or more classifications and other groupings of entities. The entities may be real world objects, like countries (grouped into groups of countries), or meta-objects, like ages (grouped into age classes).

A classification groups the entities of a register into one or more sets of mutually exclusive classes. If there is more than one set

of classes, the different sets are typically hierarchically related to each other, "level by level".

Different classifications "of the same kind" are sometimes related to each other by means of "translation keys" or "correspondences". A special case of this occurs in connection with different time versions of "the same" classification.

A grouping is a more general structure than a classification. Different groups in a grouping may overlap each other, even on the same level in a hierarchy, if there is a hierarchy, e.g. EU countries, OECD countries, Nordic countries.

Figure 2.2 illustrates some concepts needed in the description of variables and classifications. A variable (which may be derived from other variables) is associated with a value set that contains the values, which are valid values for the variable. Different variables may be associated with one and the same value set. For example, "country of birth" and "country of residence" are two different variables, but they may take their valid values from one and the same value set, "countries".
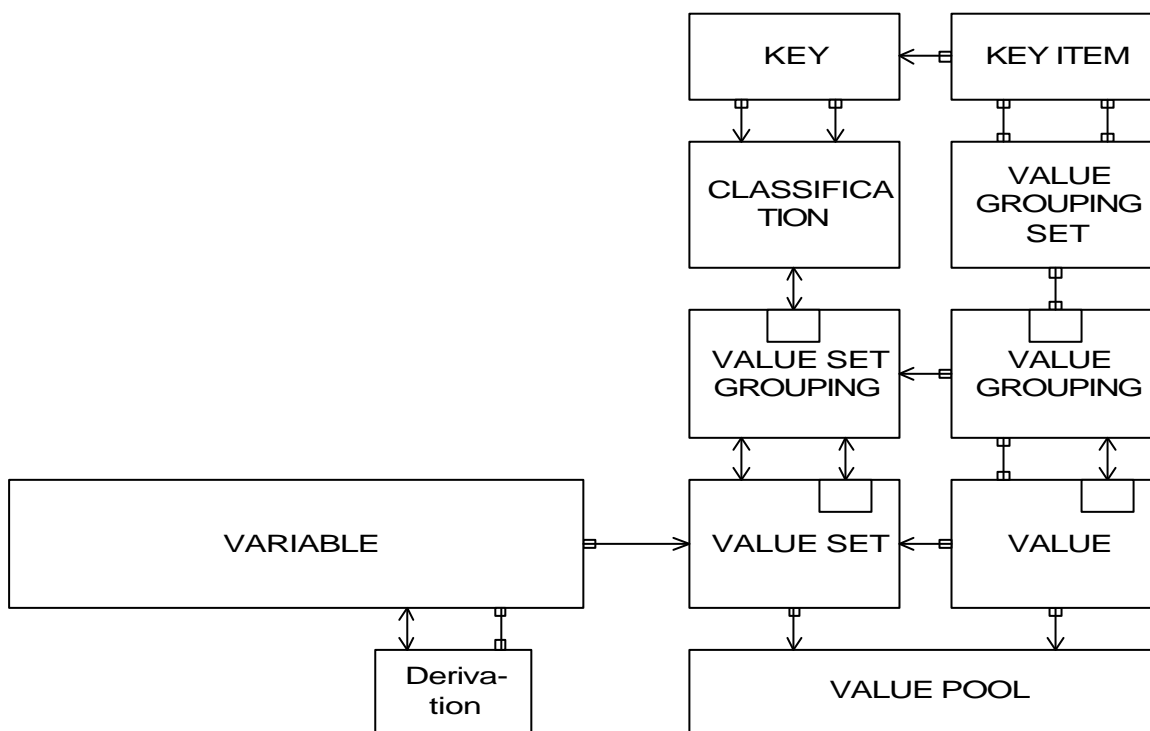
Figure 2.2 Concepts used in the modelling of a classification database, including a catalogue of variables.

Sometimes a number of value sets may be strongly overlapping in terms of values, e.g. "European countries", "EU countries", "OECD countries", "Mediterranean countries", "NAFTA countries", "NATO countries", "G7 countries"; it may be practical to view such value sets as subsets of one and the same value set, a so-called value pool.

A value set like "age groups in 5-year classes" is an example of a value set grouping. A value set grouping is itself a value set, and again different groupings of one and the same underlying value set may be regarded as subsets of one and the same value pool.

The groups in a value set grouping may or may not be overlapping. A classification is a special case of value set grouping, where the groups, called classes, overlap according to a hierarchical pattern; the hierarchy of groups consists of several levels, and each group on a certain level (except the top level) belongs to exactly one group on the next higher level.

There are often minor or (less often) major changes to a certain classification over time. As soon as there is any change made to a classification, however minor, this change results in what is formally a new classification. However, in cases of minor changes over time, it is customary to talk about different time versions of "the same" classification. Similarly, there may be different country versions and different language versions of "the same" classifications.

A key relates two different classifications (or two different versions of "the same" classification) to each other. Keys may be quite complicated. A key item describes how a class, or a set of classes, in one classification corresponds to a class, or a set of classes, in another classification. It may or may not be possible to apply the correspondences in a mechanical way, that is, it may or may not be possible to transform statistical data according to one classification in an automatic and exact way to "the same" statistical data according to another classification.

2.3   Clearing-house functions - "data warehouses"

Typically there is one common clearing-house function in a statistical organisation, serving its internal and external users. Sometimes there may be departmental clearing-house functions, serving a single department and its customers, or other special-purpose clearing-house functions, organised separately from the organisation-wide clearing-house, "the corporate database", or "the corporate data warehouse". Needless to say, in cases where there are several clearing-house functions there should be good co-operation between them in accordance with well-defined responsibilities and communication procedures, so that conflicts and undesirable redundancies and duplications of work are avoided.

From here on we shall analyse clearing-house functions as if there were only one of them per statistical organisation.

The clearing-house function of a statistical organisation should facilitate the exchange of data (and metadata) between

- survey/register functions and analysis functions (including external users),

- register functions and survey function,

- different survey functions.

An archive can be seen as a subfunction of a more general clearing-house function. Sometimes the archive function belongs to another organisation. In such cases it can be useful to maintain copies of (parts of) the archived data in the statistical organisation.

"Data warehouse" is another label for a clearing-house function that has gained popularity recently, not only in statistical organisations.

The clearing-house function should maintain the following types of statistical data:

- all final observation registers from surveys (to be kept forever);

- selected pre-final versions of observation registers (usually to be kept for a limited time);

- selected post-final versions of observation registers;

- all published statistics (to be kept forever). Ideally this component should contain all official statistics of the country, i.e. it should include official statistics produced by agencies other than the national statistical agency. A more problematic question is whether the clearing-house of a national statistical agency should also duplicate some of the official international statistics produced by international agencies. With modern Internet-based tools and standards, such duplications should ideally be unnecessary but, at least until the new techniques and standards have become perfected, there are certainly demands from some users that a national clearing-house should also contain international statistics;

- all published analyses (to be kept forever);

- metadata describing all data in the clearing-house, and the processes behind these data, to the extent and with the quality necessary for future (re)users of the data to be able to interpret and process the data correctly.

Survey/register functions, as well as analysis functions inside the statistical organisation, should deliver specified data and metadata to the clearing-house function according to specified standard formats, following specified delivery procedures.

The clearing-house function should check off data and metadata deliveries according to specified procedures, and should undertake specified actions whenever data and metadata deliveries do not occur in time or do not pass the checks.

The clearing-house function should (re)organise data and metadata in such a way that they match the needs of external and internal customers/users of the clearing-house function. The reorganisation may sometimes lead to the same data being stored redundantly in different versions, organised to meet the needs of different types of requests. As long as the data are not to be updated (in any sense other than that of new data being added), such redundancy will not cause any major problems, since storage space is relatively inexpensive.

2.4   Analytical processing systems

In a statistical organisation there will always be some analytical work performed as an integrated part of each survey. At the very least, there will (hopefully) be some production-oriented analytical work in order to produce information about the quality of the statistical outputs from the survey, and to give feed-back to more or less continuously on-going maintenance and improvements concerning the processes of the survey, as regards both quality and efficiency.

In addition to this production-oriented analytical work, there will sometimes be more user-oriented analytical functions, which are more or less independent of the survey processing systems. Of course, there will always be such functions outside the statistical organisation itself, but in many statistical organisations, not least international organisations, there are major, user-oriented analytical functions inside the organisation itself.

If a particular analytical processing system obtains all its input data from one single survey, the analytical processing system can actually be a very closely integrated part or sub-system of the survey processing system itself. In all other cases, there must be some kind of cross-survey communication. In such situations it is very natural to regard the analytical processing system as a separate system in its own right, a system with well-defined interfaces to the input-producing survey production systems. Even in the case of only one underlying survey, it would

often be advantageous to design an explicit, precisely defined interface between the survey production system and the analytical processing system.

If the statistical organisation has established a clearing-house function, a data ware-house, it is natural for analytical processing systems to demand to receive the data they need from this function, because then the data will be delivered in standardised formats accompanied by the prescribed metadata. If a clearing-house function does not exist, the analytical function will have to negotiate these things with each one of the input-delivering survey functions, and the result may be much less standardised and less predictable; if there are changes in the survey production systems, these changes may have undesirable and unexpected consequences for the analytical processing systems.

Right now, there is an emerging market for software tools aimed at very demanding analytical processing needs. So-called On-Line Analytical Processing (OLAP) tools can handle large volumes of multi-dimensional data in an efficient and very flexible way. Some statistical organisations have already started to use such tools, e.g. Oracle Express. This is excellent from an analytical point of view, but it should be noted that OLAP tools are not able to replace more standardised and general-purpose database management software tools based upon the relational data model - at least not yet. However, standard relational databases in a clearing-house function should form an excellent basis for producing OLAP databases aimed at particular user needs. Communication should be one-way from the corporate relational databases, residing on database servers, to end-user-oriented OLAP databases, residing on application servers or client computers.

CHAPTER 3

A VISION FOR THE FUTURE

Figure 3.1 illustrates a possible future information systems architecture for a statistical organisation. The architecture is based upon the four major types of statistical information systems that were analysed in Chapter 2 of this report:

- survey processing systems,
- clearing-house systems,
- registers,
- analytical processing systems.

In this proposal all clearing-house functions and register functions have been combined into one corporate data warehouse.

As was discussed earlier in this report, a typical "survey" conducted by an international organisation collects its input data and metadata from one or more statistical organisations in member countries. The observation objects of such a survey are the member countries, and the respondents are contact persons working for the national statistical office or some other agency in the member country. There are many surveys of this kind undertaken by international organisations, and they are usually operated on a repetitive basis (monthly, quarterly, and yearly). Member countries provide data in different forms; most of them deliver all data and at least some metadata in electronic form.
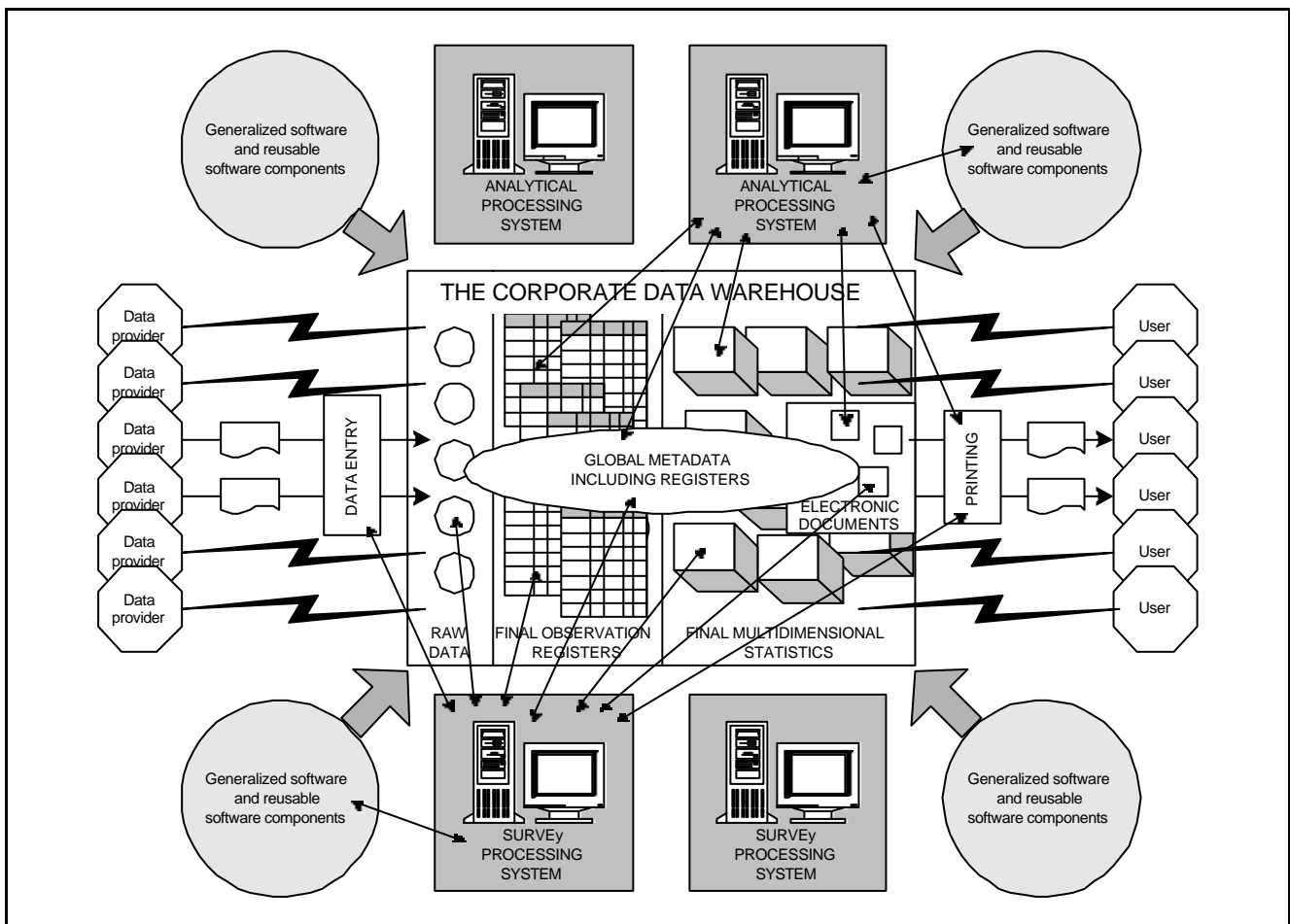

Figure 3.1 An information systems architecture for statistical organisations

Such data and metadata will go directly into the raw data compartment of the data warehouse, as illustrated in Figure 3.1. In Figure 3.1 we have included "raw data" in the corporate data warehouse. This is not completely consistent with the definition of the corporate data warehouse earlier in this report. However, there are some users of statistics, especially researchers, who insist that even the raw data should be preserved in the data warehouse, since sometimes, at least in principle, researchers may want to go back to the raw data as they were before data editing, etc. Data and metadata provided on paper have to be entered manually by a data entry process, which is managed by the survey processing system.

Even if data and metadata are delivered electronically, they may arrive in many different formats. Thus, a first step will be to standardise incoming data and metadata. This step can be avoided if member countries agree to provide data and metadata according to some international standard, e.g. the EDIFACT standard for GEneric Statistical MESsages (GESMES). It is important to note that a standard format must include standards for both data and metadata. If only the data format is covered by the standard, all metadata handling will have to be managed *ad hoc*, which typically implies iterative manual work, expensive and time-consuming.

Incoming data and metadata are transformed from raw data to final observation registers. During the transformation process, the data and metadata are checked by means of computer-aided and manual procedures; sometimes staff members have to contact the data providers. When errors and incompleteness are detected, the data and metadata are updated. At a certain stage, the data and metadata are "frozen", which means that no further updates will be made to this version of the data. The final observation register is created and stored in the final observation compartment of the data warehouse (cf. Figure 3.1); accompanying metadata are stored together with the observation register, and the global metadatabase of the data warehouse (cf. Figure 3.1) is updated.

From this point on, the observation register is available to the whole statistical organisation. Any survey processing system or analytical processing system that needs the data and its accompanying metadata can retrieve them in standard formats from the data warehouse. The survey processing system that was responsible for collecting and preparing the data and metadata will typically continue to process the final observation register into final multidimensional statistics, which will be made available to users through a variety of dissemination services (printed publications, CD-ROMs and on-line services like the Internet). All outputs are produced from the data warehouse, either directly or via a partly manual printing process. The final statistics, published by the organisation, may be stored both as standardised multidimensional tables and as print-ready electronic documents, which can be used for print-on-demand services and for HTML presentations on the Internet.

A final observation register should never be updated. If errors are detected, or if there are some other grounds for changing the contents of a final observation register, a new post-final version of the observation register should be created without deleting the original one. Similarly, if some internal user, e.g. an analytical processing system, needs to process data and metadata from a particular survey before the final observation register is ready, a separate pre-final version of the observation register should be created and stored in the data warehouse. If this preliminary use of survey data results in published statistics and/or analyses, the pre-final version of the observation register should be kept in the warehouse, even after final (and possibly post-final) version have become ready and stored in the data warehouse as well.

An analytical processing system will obtain all its input data and metadata from the corporate data warehouse, typically from the final observation register compartment and from the global metadata compartment. An analytical processing system may also use some final statistics that have already been produced and stored in the final multidimensional statistics

compartment, and that is where it will store its own final results, once they have been produced. In addition, an analytical processing system will update the global metadata compartment and will produce various end-user outputs, some of which will be stored in the electronic documents compartment of the data warehouse.

Thus, in summary, the proposed information systems architecture of a statistical organisation will consist of

- a number of survey processing systems,
- a corporate data warehouse,
- a number of analytical processing systems.

The data warehouse will contain "compartments" for

- raw data and metadata,
- final observation registers,
- final multidimensional statistics,
- electronic documents,
- global metadata, including registers.

The architecture contains "residual", partly manual functions for data entry and printing, but these functions will be "driven" by the data/metadata structures in the data warehouse, and they will be controlled as integrated parts of the survey processing systems and/or analytical processing systems by the people and organisational units responsible for those systems.

3.1   Survey processing systems

In principle, all survey processing systems of statistical organisations, national and international, run through the same kind of life cycle as was described in section 2.1:

1. Survey planning
1.1 Specify major users and purposes of the survey
1.2 Specify major outputs from the survey
1.3 Specify major inputs to the survey
1.4 Specify main procedures for obtaining inputs and transforming them into outputs

2. Survey operation
2.1 Frame creation
2.2 Sampling
2.3 Measurement
2.4 Data preparation
2.5 Observation register creation
2.6 Estimation and analysis
2.7 Presentation and dissemination

3. Survey evaluation
3.1 Check that specified outputs from the survey have been delivered properly
3.2 Check that production-oriented metadata have been properly collected and stored
3.3 Organise feedback information from users
3.4 Evaluate feedback information from the production process and from the users

There are some important differences between a survey conducted by an international organisation and a survey carried out by a national statistical office:

- Frames and sampling procedures do not play important roles in international surveys;

- Estimation and analysis is more complex in the international environment than in a national statistical office.

3.2   Data warehouse - including registers

This proposal identifies and defines five compartments of the future corporate data warehouse function of a statistical organisation:

- raw data and metadata;
- final observation registers;
- final multidimensional statistics;
- electronic documents;
- global metadata, including registers.

Data and metadata in the raw data compartment will sometimes be in standardised form, e.g. the GESMES format, sometimes not. There should be generalised software supporting the standardisation of data and metadata, once they have arrived. From then on, there should be generalised software tools supporting all important processes and sub-processes in survey

processing systems and analytical processing systems. Data and metadata in the data warehouse constitute well-defined standard interfaces between systems and between processes within systems.

In the context of a national statistical organisation, the final observation registers typically contain microdata, and they are typically stored as relational tables (in the sense of the relational data model) or as so-called flat files, which are logically more or less equivalent to relational tables. An international statistical organisation will not normally handle microdata from member countries. On the other hand, the macrodata provided by member countries to international organisations may formally be handled in very much the same way as microdata are handled by a national statistical organisation; in effect, if we adopt another perspective, the macrodata provided by member countries actually are microdata, viz. if we regard the individual countries as observation object on another micro-level.

Hence it is proposed that the input data of an international statistical organisation should be handled by the same types of data structures and the same types of software tools that are used for the management and processing of microdata in national statistical organisations. Relational database management systems and software tools compatible with the relational data model are today *de facto* standards for such tasks, and they are likely to remain so for the foreseeable future (3-5 years).

Thus the final observation registers in the data warehouse of all kinds of statistical organisations, national and international, should be stored in relational tables in relational databases.

Generalised software tools can be used for (further) aggregating and analysing the statistical data in the final observation registers into final, multidimensional statistics to be presented and disseminated by the statistical organisation.

Whereas the relational data model was immediately accepted as suitable for the management of statistical microdata, it has been debated among statisticians whether it is also suitable for the management of statistical macrodata. It has been argued that the relational data model does not do justice to the multidimensional character of macrodata, and that the management of multidimensional data becomes inefficient if these data are mapped into "flat" relational tables.

Thanks to the dramatic price/performance development of information technology, the efficiency problems have increasingly become "non-problems". Today there are no problems with the handling of all statistical data, including multidimensional macrodata, of most statistical organisations by means of standard commercial relational software on standard, inexpensive PC-based equipment. As an example, the total volume of OECD statistical data, to be stored in a corporate data warehouse, has been estimated at 10-15 GB. Even if this should increase by 10 times, and is stored in a less hardware-optimised way, it could still be easily accommodated on a standard PC-based database server.

There is an emerging market for new kinds of generalised software products for very efficient and flexible management of large volumes of multidimensional data. These software tools are sometimes called tools for On-Line Analytic Processing (OLAP) as opposed to tools for On-Line Transaction Processing (OLTP). Whereas OLTP tools are optimised for efficient handling of update-oriented processes (like those processes that occur during the data collection and preparation phases of a statistical survey processing system), OLAP tools assume that data will not be updated, which means that the database can be optimised for the processing of complex *ad hoc* queries (e.g. unplanned statistical queries versus a large statistical database).

OLAP tools have a great potential for statistical organisations, but for the foreseeable future (3-5 years) they can only be a complement, not a replacement, for relational

software, because they are not suitable for the update-oriented parts of statistics production. Thus an OLAP tool cannot be the *basic* tool for database management in the statistical organisation's information system infrastructure. On the other hand, it is easy to produce an OLAP database from a relational database, so internal or external users of data from the data warehouse can easily "add on" OLAP tools for analytic processing of the data, if they so please.

According to the proposed architecture, statistical end products can be stored in two different ways which are not mutually exclusive:

- as multidimensional data, stored in relational tables, with accompanying metadata, stored partly in relational tables, partly in text databases;

- as electronic documents, in SGML format, ready for printing on demand, Internet HTML publishing and other "one-way" dissemination forms.

The first alternative has the advantage that the end-product can be reused by the same processes and software tools as those by which it was itself produced, that is, it can be reused as an input to other statistics production and analysis processes within or outside the statistical organisation.

The second alternative is attractive as a cost- and time-efficient option to traditional publishing procedures, while still maintaining very high-quality printing possibilities.

OLAP databases and other special-format products can be viewed as a third category of future statistical end products.

The global metadata component of the proposed architecture is shared by all other components. It contains both structured and more or less unstructured metadata.

Structured data can be stored in relational tables and can be handled by the same relational software as the data that they describe. This

way of handling metadata is suitable when the metadata are to be used for driving computerised processes, since the software controlling such processes are then able to process data and metadata in a completely integrated and automated way.

Metadata about variables, value sets and classifications are good examples of structured metadata that should be stored in relational tables and managed by standard relational software products.

To a limited extent, relational software is able to handle less structured text data as well, but it is usually more efficient and user-friendly to manage less structured metadata by means of generalised text management tools, e.g. word processors and free-text search systems.

Large parts of the documentation of a statistical survey can best be treated as less structured free-text metadata. On the one hand, this lessens the burden on the metadata provider, and on the other hand it facilitates very fast and flexible searches for data and metadata.

Figure 3.2 visualises a data model for the metadata of a statistical data warehouse, a so-called metadata model. The model consists of three major parts:

- a microdata and survey-oriented part, including process-oriented metadata;

- a macrodata oriented part, including user-oriented quality metadata;

- a classification oriented part, including metadata about value sets and classifications.

In the case of an international statistical organisation, the microdata and survey-oriented part of the model primarily describes the properties of the national surveys underlying the corresponding "surveys" conducted by the international organisation; more or less unstructured textual metadata will describe how the typical survey procedures (as visualised
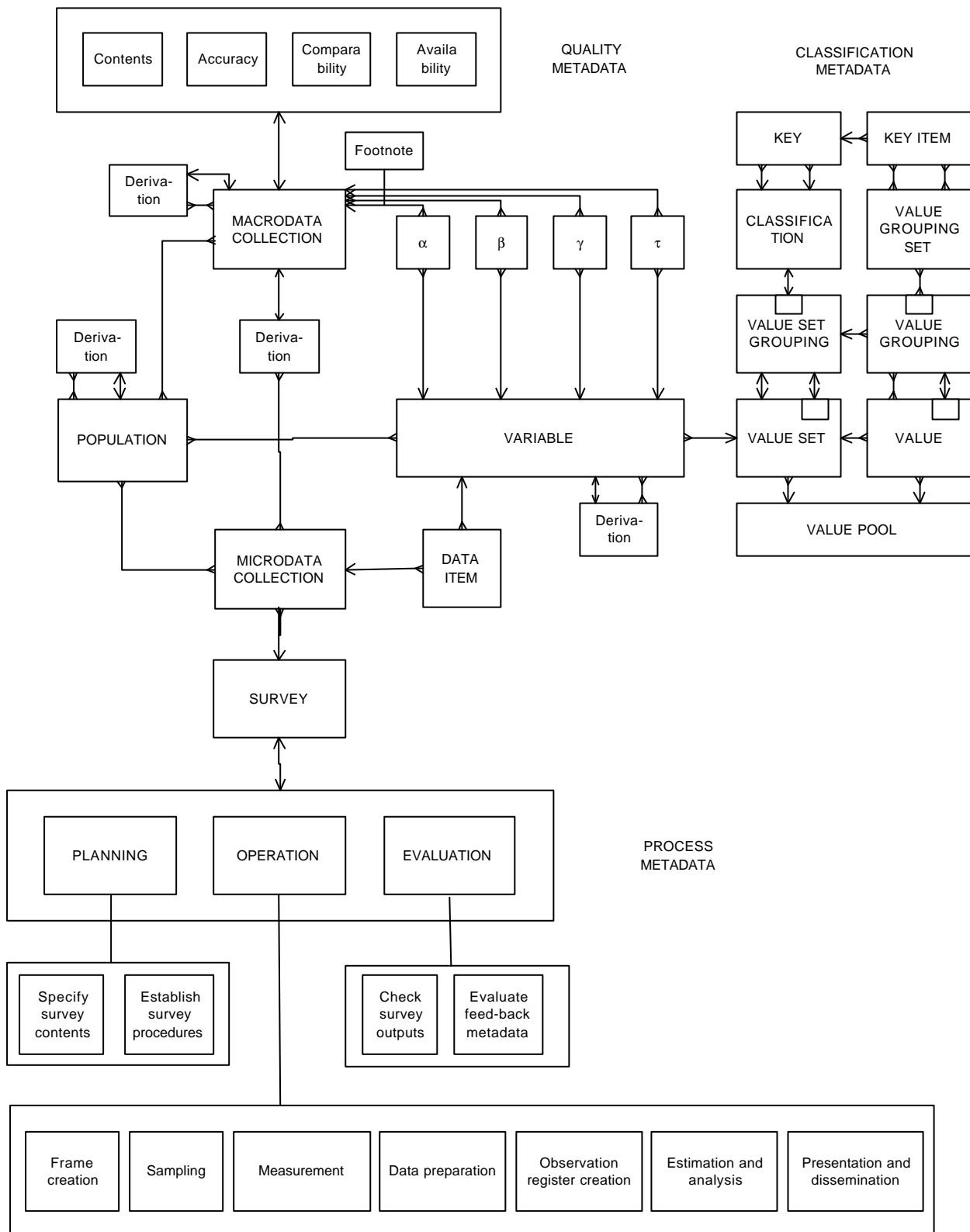
Figure 3.2 Metadata model for a statistical data warehouse

earlier in Figure 2.1) are designed and carried out in each one of the member countries of the organisation; some structured metadata are collected concerning the variables and populations referred to by the data in the microdata collections produced by member countries.

The macrodata-oriented part of the metadata model describes all macrodata collections that are either received or produced by the statistical organisation. A macrodata collection is a set of macrodata that contains estimated values of statistical characteristics concerning (at least) one population of objects; cf. the definitions in section 1.1.2 of this report. According to the same definitions, a statistical characteristic is a measure that summarises the values of (at least) one variable of the objects in a population. A macrodata collection may always be thought of as having been derived from an underlying microdata collection containing observations of the variable(s) summarised by the statistical characteristic. As a matter of fact, this is often not only a way of thinking, it is also the way the macrodata have actually been produced, possibly in several steps, where some of the steps have occurred in respondent organisations. The careful reader may observe that the data model in Figure 3.2 is a little more general than the data model assumed in Figure 3.1 for a typical data warehouse of a statistical agency. For example, in Figure 3.1 all microdata collections (except raw data collections) are final observation registers. The data model in Figure 3.2 does not assume that a microdata collection is necessarily a final observation register.

A more or less formal description of a macrodata collection can be structured along four major dimensions:

α: the alfa dimension, the dimension of population and scope;

β: the beta dimension, the dimension of measurement and summation;

γ: the gamma dimension, the dimension of classification;

τ: the tau dimension, the dimension of time.

Figure 3.3 illustrates this by means of a simple example, taken from the OECD Main Economic Indicators (MEI). The example is analysed both from the perspective of an international statistical organisation, in this case the OECD (interpretations 1a, 1b, 1c), and from the perspective of a national statistical office (interpretations 2a, 2b, 2c).

Interpretations 1a and 2a use concepts from the definition of a statistical characteristic. According to the OECD perspective (interpretation 1a), the observation object is Canada, an OECD country. According to the member country perspective (interpretation 2a), the observation objects are crude steel producers in Canada. In both perspectives the observation variable is "production of crude steel during July 1996".

Interpretations 1b and 2b emphasise certain structural aspects of the analysed statistical message. Firstly, time is parameterised by explicitly regarding "July 1996" as a value on the time scale "month". Secondly, it is explicitly recognised that "crude steel" is a value in the "commodity" value set. The latter structure is moved from the observation variable part of the message to the population part. According to this view, the population consists of <country, commodity> pairs (interpretation 1b) and <commodity producer, commodity> pairs (interpretation 2b), respectively.

The $\alpha\beta\gamma\tau$ scheme of analysis used in interpretations 1c and 2c makes the structure even more explicit and visible, and it stresses that the statistical message can be seen as the content of a very general, multidimensional table, spanned by the γ-variables; in the example analysed, it is a three-dimensional table, spanned by the variables "country", "commodity", "month". A formal time parameter, τ, is used to indicate time aspects in a systematic way, wherever relevant. In the analysed example τ occurs both as a point of time (indicated by the key word "<u>at</u>" before the time parameter) and as a time interval (indicated by the key word "<u>during</u>" before the time parameter). In a more complex example different functions of τ could have occurred in

"Canada's production of crude steel during July 1996 was 1126 thousands of tonnes."
OECD Main Economic Indicators October 1996, p 48.

Interpretation 1a: observation message, OECD perspective

observation population: OECD countries July 1996
observation object: Canada
observation variable: production of crude steel during July 1996
observation value set: thousands of tonnes
observation value: 1126

Interpretation 1b: More structured observation message, factorisation and parameterisation

observation population: <OECD countries (month), commodity>
observation object: <Canada (July 1996), crude steel>
observation variable: production (month = July 1996)
observation value set: thousands of tonnes
observation value: 1126

*Interpretation 1c: multi-dimensional table, **abgt**-structure*

α: <u>for</u> <OECD country (<u>at</u> τ), commodity>
β: <u>give</u> <u>sum</u> (production (<u>during</u> τ) in thousands of tonnes)
γ: <u>by</u> country, commodity, τ
τ: <u>where</u> τ = month

Interpretation 2a: statistical message, member country perspective

observation population: crude steel producers in Canada, July 1996
observation variable: production of crude steel during July 1996
observation value set: thousands of tonnes
measure: sum
estimated value: 1126

Interpretation 2b: More structured statistical message, factorisation and parameterisation

observation population: <commodity producers in OECD countries (month), commodity>
observation object: <Canada (July 1996), crude steel>
observation variable: production (month = July 1996)
observation value set: thousands of tonnes
measure: sum
estimated value: 1126

Interpretation 2c: multi-dimensional table, country perspective

α: <u>for</u> <commodity producer in OECD country (<u>at</u> τ), commodity>
β: <u>give</u> <u>sum</u> (production (<u>during</u> τ) in thousands of tonnes)
γ: <u>by</u> country, commodity, τ
τ: <u>where</u> τ = month

Figure 3.3   Different structuring schemes for analysing statistical data from the OECD Main Economic Indicators

different parts of the structure; example: "increase in income between year τ and year $\tau+1$".

The quality metadata of the macrodata part of the metadata model should in principle be derivable from the metadata concerning the underlying microdata collections and microdata-producing surveys in the microdata part of the metadata model. In practice, the derivation may be rather complex, and it may require considerable human expertise and judgement.

3.3   Analytical processing systems

An analytical system uses data and metadata from one or more surveys, sometimes in combination with data and metadata from other sources. According to the architecture proposed here, the analytical processing systems should be able to obtain the data and metadata that they need from the corporate data warehouse. Published results and other end-products from analytical processing systems should be stored in the data warehouse.

Sometimes the results of an analysis may indicate that there is something wrong, or at least questionable, about the underlying data. Such findings may even lead to contacts with survey respondents in member countries, and these contacts again may result in revised data and/or metadata. It is important to have well-defined procedures for handling such revisions. The main alternatives are:

- If the final observation register of the survey has not yet been created: to take the revised data/metadata into account, when creating the final observation register;

- If the final observation register has already been created: to create a new, post-final observation register, while keeping the already existing final observation register;

- Making no changes to the underlying observation registers, but making some kind of comment and/or footnote in connection with the presentation of the results of the analysis.

CHAPTER 4

TECHNICAL ASPECTS OF THE PROPOSED ARCHITECTURE

This chapter of the report starts from some basic ideas on how human beings have learnt to cope with the phenomena that are so complex that all relevant aspects of them cannot be fully grasped by the human brain at one and the same time. Such complex phenomena are sometimes called imperceivable systems, and one approach to the understanding, control and (in the case of man-made systems) design and maintenance of such systems is called the systems approach.

We will then move on to discuss how to apply the systems approach to the design and management of the proposed architecture for the information systems of a national or an international statistical organisation.

4.1 The systems approach

The systems approach is a general human approach for describing, analysing, and controlling complex phenomena, i.e. phenomena that consist of a large number of related aspects, which the human brain is unable to fully appreciate at once. The systems approach is an attempt to manage complexity by combining precise analyses of details with a good understanding of the whole.

4.1.1 Basic ideas

Some basic propositions of the systems approach are:

- a complex phenomenon can be conceptualised as a system, a so-called imperceivable system, since it cannot be fully understood by a single mental act;

- a system consists of parts;

- a part of a system is in itself another system, a subsystem of the former system;

- any system, even the whole phenomenon first considered, is a part of a wider system, a supersystem, or environment, of the former system;

- the parts of a system are related to each other, and to the system as a whole, and the system is related to its parts as well as to other systems in its environment.

There are natural systems and designed systems. A natural system exists as a phenomenon independently of human beings, although the conceptualisation of the phenomenon as a system (adhering to principles like those mentioned above) is, of course, a product of human thinking, which does not necessarily exist independently of human beings.

A designed system is a phenomenon that in itself is the result of human thinking; it would not exist without conscious and explicit human effort. Computers and computerised information systems are examples of designed systems.

It is a philosophical question whether natural systems serve certain purposes, or whether they just exist without any particular purpose. Designed systems, on the other hand, are typically designed to serve one or more purposes. For such systems, it is clearly meaningful to talk about efficiency; an efficient system is a system which serves its purposes in an efficient way.

There are at least two aspects of the efficiency of a system, represented by the following two questions:

- To what extent does the system fulfil its purposes?

- How many resources (of different kinds) does the system use?

The first efficiency aspect can be called "goal fulfilment efficiency", and the second efficiency aspect can be called "resource efficiency".

When the two aspects of efficiency are combined, one talks about the overall cost/performance, or cost/benefits, of the system.

The resources used by a system may be of many different kinds: material resources, energy, human capacity (physical and/or brain capacity), time, etc.

In order to simplify the evaluation of the efficiency of a certain system, one often tries to measure all kinds of goal fulfilments as well as all kinds of resource usages in monetary terms. If the different kinds of goal fulfilments and the different kinds of resource usages occur at different times during the life cycle of the system (including the design phase of the system), which they typically do, one tries to come to a "correct" monetary evaluation by discounting the monetary effects to one common point of time.

4.1.2   Are there efficient system architectures?

A designed system can typically be designed in many different ways, representing different patterns of goal fulfilment and resource usages over the life cycle of the designed system. It is an interesting question, whether there are certain general features that make a certain design, a certain system architecture, more efficient than alternative designs, alternative system architectures.

A basic assumption behind the systems approach is that it is efficient to conceptualise complex phenomena as systems, consisting of related subsystems, etc. Furthermore, the systems approach seems to prescribe that a complex phenomenon/system should be conceptualised in terms of subsystems in such a way that

- there are few and simple relationships between subsystems - "simple standard interfaces",

- there are few and simple parts on the lowest level of the system/subsystem hierarchy - "simple standard components".

4.2   The systems approach and statistical information systems

It is a rather straightforward task to apply the systems approach to computerised information systems in general, and to statistical information systems in particular. Obviously, such a system is a designed system. The purpose of the system is to provide certain information and information-related services to its users. Statistical information systems in particular are often multi-purpose systems in the sense that they should provide information to many different kinds of users, with different and sometimes contradictory needs, now and in the future. Thus, flexibility is a particularly important consideration for statistical information systems.

We have established earlier in this report that a statistical organisation typically runs a relatively large number of information system applications; a typical national or international statistical organisation may run hundreds of applications. However, many of these applications are rather similar in the sense that they perform a limited number of functions, which are characteristic of information systems supporting statistical surveys:

- survey planning functions and subfunctions;

- survey operation function and subfunctions;

- survey evaluation functions and subfunctions.

In addition to the survey processing systems, we have noted that there are registers,

data warehouses and analytical processing systems, and these systems have rather typical functional characteristics, too, in the statistical environment.

It is certainly possible to exploit the advantages of this regularity in types and functions of statistical information systems by applying the systems approach. In particular, when we design the information systems of a statistical organisation, we should establish a systems architecture which supports the possibilities to

- use standard interfaces for communication of data and control between systems and between subsystems within systems,

- use standard components for the realisation of systems and subsystems.

When we move from the design to the implementation phase of the information system life cycle, we can fully exploit the inherent simplicity, robustness and flexibility of such a systems architecture.

The main advantages of this approach are that the approach

- makes it easier for human beings to understand and reflect about the system,

- increases possibilities to reuse existing components rather than developing new ones,

- reduces construction work, construction time, and construction errors,

- supports flexibility, making it easier to add functionality to the system in the future,

- simplifies maintenance work and makes it easier to adapt to changing technology.

Like other computerised information systems, statistical information systems are implemented by means of hardware, software and data components. We shall now discuss how the systems approach can be applied to each one of these component types.

4.2.1   Hardware components

Today it is rather easy for a statistical organisation to choose standard hardware components. The IBM compatible PC has since long become a *de facto* standard hardware component for statistical organisations and, to an even greater degree, for the users and customers of statistical organisations. Until recently statistical organisation often judged it to be necessary to maintain other hardware components as well, e.g. mainframes and minis, but today most statistical organisations would be better off abandoning these non-standard hardware components all together. The capacity of a standard PC is enough for all, or almost all, statistical needs, as well as for all the administrative non-statistical needs of most statistical organisations. It should be stressed that the argument presented here in favour of a uniform PC-based platform is not primarily based upon technical considerations but on the economical and budgetary realities facing most statistical agencies today. It is simply too expensive for most agencies to maintain the competence for more than one technical platform. However, for agencies that are not (yet) facing such economical restrictions, it may be argued that a multi-platform environment, including mainframes, UNIX boxes, PCs, and maybe even Macs, enable solutions that are more elegant and more efficient from a technical point of view.

4.2.2   Software components

A software system consists of software components and interacts with hardware and data components. In the case of statistical information systems there should be a correspondence on a certain level in the systems architecture between

- on the one hand: the functions and subfunctions of survey processing systems (and other types of statistical applications);

- on the other hand: software components of the software applications supporting the

statistical applications and their functions and subfunctions.

A software system used for implementing an information system application is called a software application. In the information systems architecture of a statistical organisation, there should be a correspondence between software applications and statistical applications. When establishing this correspondence, one should try to reuse generalised software systems: commercial packages or, presumably less often, in-house developed. An alternative is to tailor an application on the basis of reusable, commercial or home-grown software components. The two alternatives may be combined.

When a statistical function or subfunction is analysed, at some stage one reaches a level where the software components need not necessarily be tailored to the needs of statistical applications. Instead one may use general-purpose software components. Again, one should try to exploit reusable, standard software rather than tailoring application-specific software components. Nowadays, this "general-purpose level" may appear rather high up in the systems architecture of a statistical application. For example, a word-processing system or a spreadsheet program may be a high-level software component of this kind. So-called software objects are examples of lower-level reusable software components. An application software developer may look for reusable software components on the software market as well as in-house, in software component libraries, which should be organised by application-developing organisations, e.g. statistical organisations. Co-operation between statistical organisations may also be an alternative for developing useful software component libraries.

The interaction between application software and hardware takes place through so-called systems software, such as operating systems. A certain combination of hardware and systems software is often identified as a particular (hardware/software) platform, e.g.

- IBM compatible PC in combination with Microsoft Windows NT,

- mini-computer in combination with a UNIX dialect for that mini-computer.

From an economical efficiency point of view it is most advantageous if an organisation can avoid having more than one hardware/software platform for its information system infrastructure. Every additional platform adds significantly to the complexity of the technical infrastructure of the organisation and necessitates a considerable extra amount of technically competent staff.

A note on reusable software components

A reusable software component is a software component that can be used, without being changed, in several different software systems. A software component may consist of other software components. A whole software system, e.g. a word processing system, may be a software component of another software system, e.g. a statistical information system.

A software application will consist of reusable and non-reusable software components. Reusable software components are sometimes called generalised software components, or standard software components, whereas non-reusable software components are sometimes called tailor-made software components. Reusable software components can be developed by professional software producers, and marketed by professional software vendors. However, reusable software components can also be produced and shared internally within any kind of organisation that develops software applications, e.g. a statistical organisation.

A software component is associated with

- one specification;
- (at least) one (source-code) implementation;
- (at least) one executable (run-time) unit.

A reusable software component should have a precise specification, defining the function of the software component, as well as its external

interface in terms of inputs and outputs. A software component is reusable, only if it can be plugged into other software systems, exactly as it is, without any changes or modifications. The user of a reusable software component should not need to take any responsibility for the maintenance of a reusable component. If an existing software component is modified and then plugged into an application, this may speed up the development process in comparison with tailoring a completely new software component from scratch, but strictly speaking this is not reusage; the application developer will have to take full responsibility for the maintenance of the modified component. However, modification of a (reusable) software component may result in another reusable software component.

4.2.3 Data components

The data components of an information system are stored either as physically integrated parts of the application software system or as separate files or databases. One of the major contributions of the database concept and the ideas of database-oriented information systems development was the introduction of the concept of (different degrees of) program/data independence. Program/data independence means that the software and data components of an information system may be developed and maintained relatively independently of each other. A modification of the contents, structure or storage of data should not necessitate modifications of programs using the data. On the other hand, it should be possible to modify or add software components without having to redefine data components.

In order to facilitate the realisation of the concept of program/data independence, a new type of interface was developed: the interface between application programs and databases. A special type of systems software, called database management software, was developed in order to provide the program/data interface functionality. The four basic functions of a database management system are to

- add,
- retrieve,
- update, and
- delete

specified data. In order to be able to specify the data upon which these operations should be performed, there must be a data model describing the data stored in the database, and the database management system must be able to interact with the database in terms of this data model.

Soon after the database concept had been introduced in the beginning of the 1970's several competing proposals for standardised database interfaces were presented. IBM tried to establish a commercial *de facto* standard, based upon a hierarchical data model, reflecting how data were handled by IBM systems software. Non-IBM vendors and some users tried to agree upon an international standard, based upon a network-oriented data model, developed by the CODASYL Data Base Task Group. University researchers favoured the relational data model, which was aimed at being closer to how users conceptualised data in databases, and less dependent on how data were actually stored on different physical media. So-called conceptual data models, like the Infological Approach and the Entity Relationship model were proposed as well, and they tried to be even closer to users' conceptualisations of information.

There was, and is, a need to reconcile

- conceptual models, reflecting the "objective" abstract information contents and end-users' different views of data,

- external data models, reflecting the different views of data taken by application programs and application programmers, and

- internal data models, reflecting how data are physically stored.

In order to accomplish the necessary reconciliation between these different types of models, so-called multi-schema architectures were developed for database interfaces, supporting mappings between different schema levels (e.g. the external level, the conceptual

level, and the internal level), thus permitting the same physical data to be viewed and used in many different ways within one and the same information system infrastructure. The ANSI/SPARC proposal was a pioneering three-schema architecture.

Today the relational data model is the *de facto* standard for a wide range of commercial database management systems and database-related software products. The Structured Query Language (SQL) is the equally widely accepted interface between relational database management systems and database-related software products.
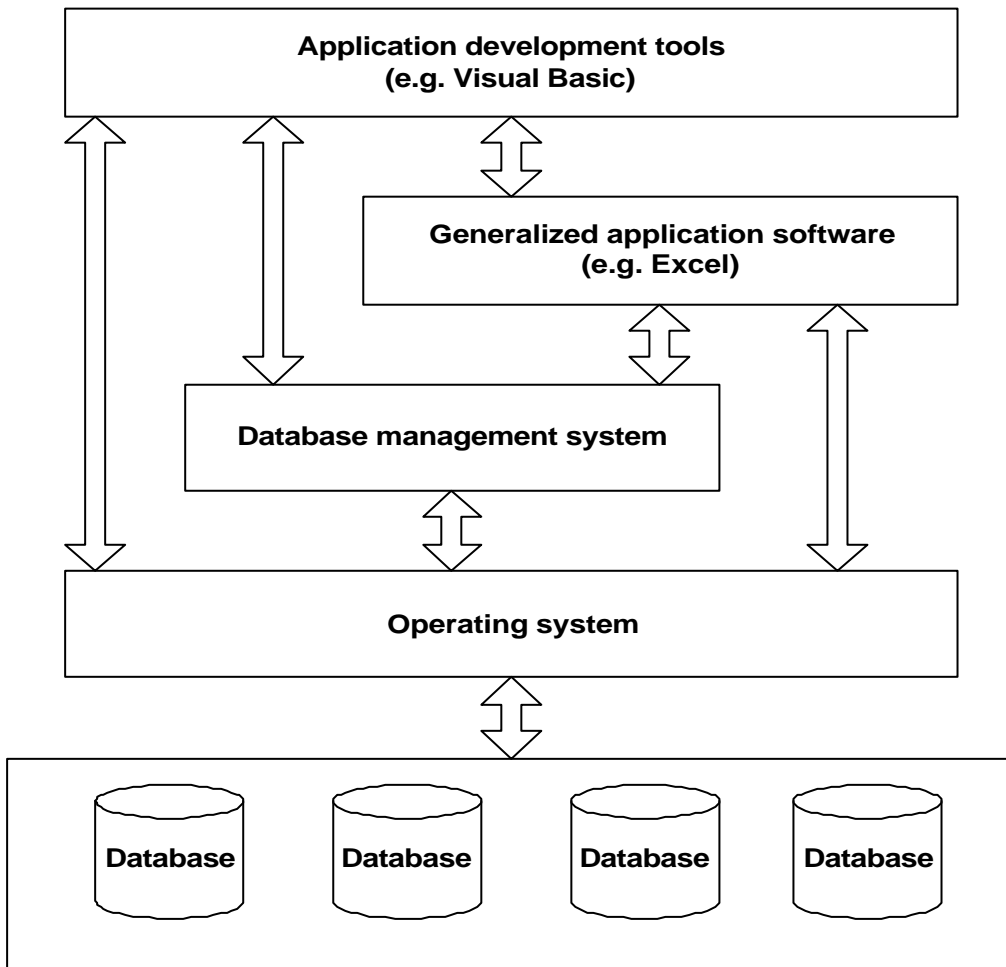
4.2.4 *Interaction between software components and data components*

Figure 4.1 summarises the different kinds of interactions that occur in a computerised information system between different kinds of software components, as well as between software components and hardware components. Each interaction type requires a well-defined interface.

4.3 A multi-tier network architecture for statistical organisations

This section will lead to a proposal for a network-based information systems architecture that balances the needs for centralisation and decentralisation in a modern statistical organisation. In order to illustrate the needs for such a balance, we shall start with a brief review of computer history which, although short, displays examples of both extreme centralisation and extreme decentralisation.


Figure 4.1 Interaction between different types of software components and data in the databases

4.3.1 Historical starting-point: mainframe-based centralisation

Originally, computerised information systems were completely centralised in all respects. All computerised activities were organised around the mainframe computer, including input/output processes, data storage, and data management and computation processes. Even application system developers and application programmers belonged to a centralised organisation around the computer; inputs in the form of card decks were delivered to a desk outside the "closed shop" computer room, and output listings were picked up later, often much later, in the same place. As in some other centrally planned systems, queues were common.

4.3.2 Top-down distribution of functions: dumb terminals

With the introduction of terminals in the early 1970's some computerised activities were moved a little bit closer to where they naturally belonged in the organisation. End-users could deliver some input data to the computerised information systems through the terminals, and they could get answers to simple, structured questions through terminal-based menu systems and non-procedural command languages. Programmers could enter and edit their source code and control statements via the terminals. But all intelligence remained in the mainframe. The terminals were all so-called dumb terminals.
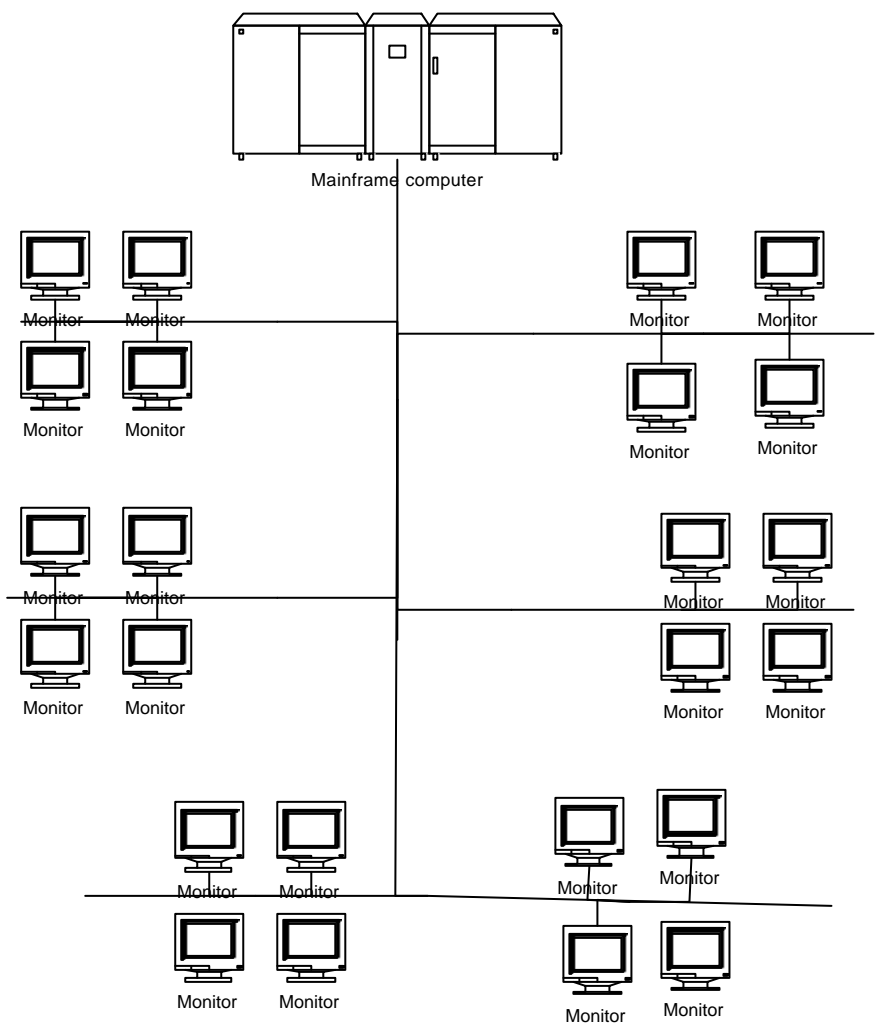

Figure 4.2 Mainframe-based architecture

4.3.3 *Extreme decentralisation: personal microcomputers*

With the personal microcomputers introduced in the early 1980's the organisation of some computerised systems went to the opposite extreme, all activities being done locally. The computerised processes that now became organised in better harmony with other aspects of the organisation were far from core business processes, but the decentralised computerisation of word processing, spreadsheet calculations and other basic administrative support functions still had a dramatic effect on the impact and organisation of computerised information systems. For the first time, virtually all employees could be in control of some computerised processes. More and more users demanded to be in control of computerised systems rather than being controlled by the systems. The change was strongly supported by new employees from the universities, who had

got used to the new, active way of using computers, and they expected to continue to work in this way when they entered business life.

4.3.4 *Bottom-up co-operation and resource-sharing: networks*

A completely decentralised organisation of computerised activities has its drawbacks. Most business activities have some natural needs for communication with other activities, controlled by other persons in other parts of the organisation, or even outside the organisation. For example, in a statistical office an important part of the overall information potential stems from the statistical organisation's ability to combine data from different sources, e.g. from different registers and surveys operated by different parts of the statistical organisation, and even from sources outside the statistical organisation. In order to exploit these possibilities fully, there are not only needs for
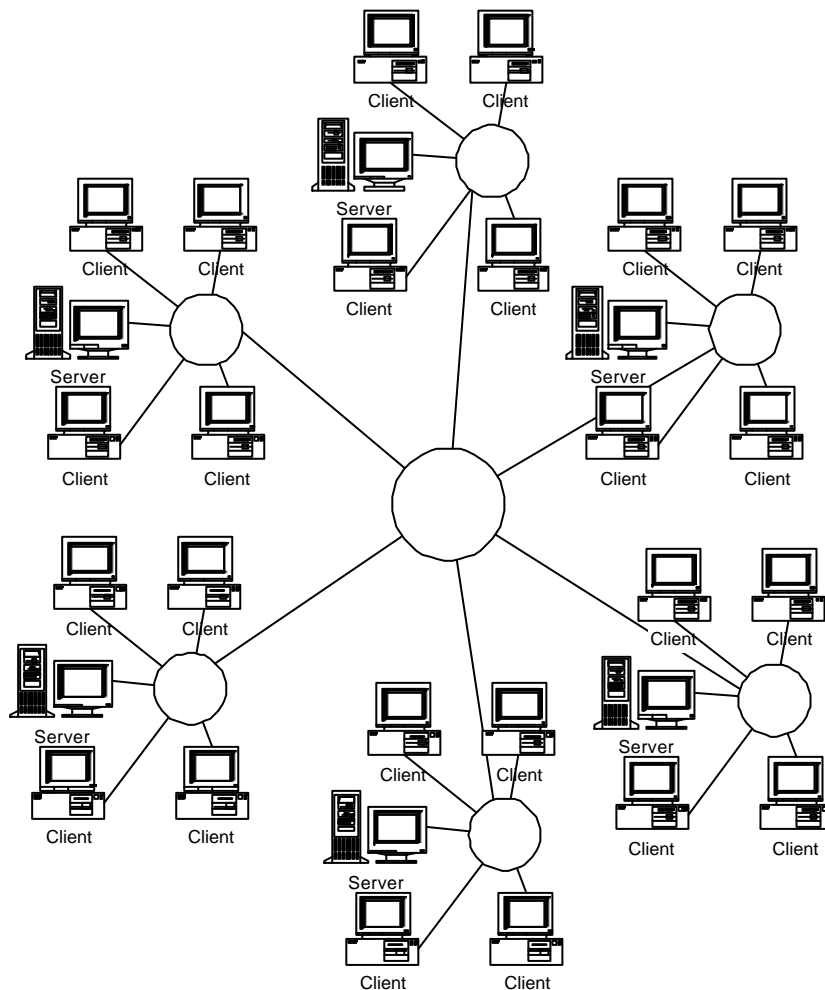
Figure 4.3 Communication and resource-sharing in a network

communication between different information systems; there are also needs for technical and contents-oriented co-ordination, which requires some degree of centralised control. However, it should be noted that this kind of centralisation need does not emanate from shortcomings or costs of the technology used, but from the nature of the business as such. For example, the production of certain types of statistics, which are demanded by the users, require definitions of objects, populations and variables to be co-ordinated and even standardised, so that different kinds of observations, made by different surveys and information systems, can be combined and used for many different purposes.

As a consequence of obvious, and at first rather simple co-ordination needs, the personal microcomputers of an organisation very soon became linked to each other through local area networks (LAN) and wide area networks (WAN). File servers and printer servers made it possible for several users to share the same data and printing resources, respectively; cf. Figure 4.3.

It should be noted that resource sharing and co-ordination in a network of personal microcomputers is rather often, and maybe most efficiently, driven in a "bottom up" fashion, starting from natural needs of persons and local organisations involved in the network. This is in contrast to mainframe-based distribution of resources, which is typically driven by the computer department in a "top down" manner. It is often argued by computer departments that a certain degree of centralised control, applied "top down" is necessary for the proper functioning of the information system infrastructure of an organisation. The recent success story of Internet is a strong argument in the opposite direction. The Internet is a striking example of a truly decentralised network, where almost all co-ordination and control is applied "bottom up", and nobody could claim that this network as a world-wide information infrastructure is less complex than any company-internal information system.

Thus there are strong arguments in favour of a development strategy where the individual applications voluntarily develop their co-ordination to a large extent, without much centralised control. As in the Internet, there are obviously needs for certain standards, so that efficient communication and resource sharing can take place. It is often an advantage if these standards can be voluntarily agreed upon as *de facto* standards, but sometimes formal decision-making may be needed in an organisation to speed up the process, if nothing else. But it should never be forgotten that many successful standards have become successful in part because they are accepted - and abandoned - as the result of the free will of participants in a network.

In the framework of an organisation, a local actor may often find it advantageous, at least temporarily, taking competence and prices into account, to give up some local authority even over some tasks that need not be co-ordinated or centralised because of inherent features of the business. When giving up authority for other than business-inherent reasons, the person or organisation who gives up authority should make sure that the process is reversible, so that the delegation of authority remains a delegation and does not become an abdication, and so that the delegation can be taken back, if the assumptions on which it is based, e.g. cost relationships, change at some stage in the future.

4.3.5 *Client/server server architecture*

A conclusion from the discussion so far is that the resources and tasks in a computerised information system should be organised in such a way that this organisation is in full harmony with the business activities of the organisation. If there is strong local control over a certain business activity, this control should cover the corresponding functions of the supporting information systems as well. To the extent that there are business-related needs for resource sharing and co-ordination, this extends to the corresponding functions of the supporting information systems as well.

The client/server (c/s) architecture (cf. Figure 4.4) is an attempt to translate this philosophy into design principles for a network of co-operating nodes. In the original version of the c/s architecture the network nodes were classified into two categories:

- client nodes, with typically one computer per person using the network;

- server nodes, with typically one computer per function, serving a group of users.

Some functions which are very often shared by a group of users are printing, data storage and communication functions, e.g. e-mail, bulletin boards, and conferences. The group being served by one server may be anything from a small, natural working group in the organisation, up to a department, a division, the whole organisation, or an even larger community. For example, a printing function is typically shared by a small working group, whereas the database service of a statistical organisation may be shared by a large community of users inside and outside the statistical organisation.
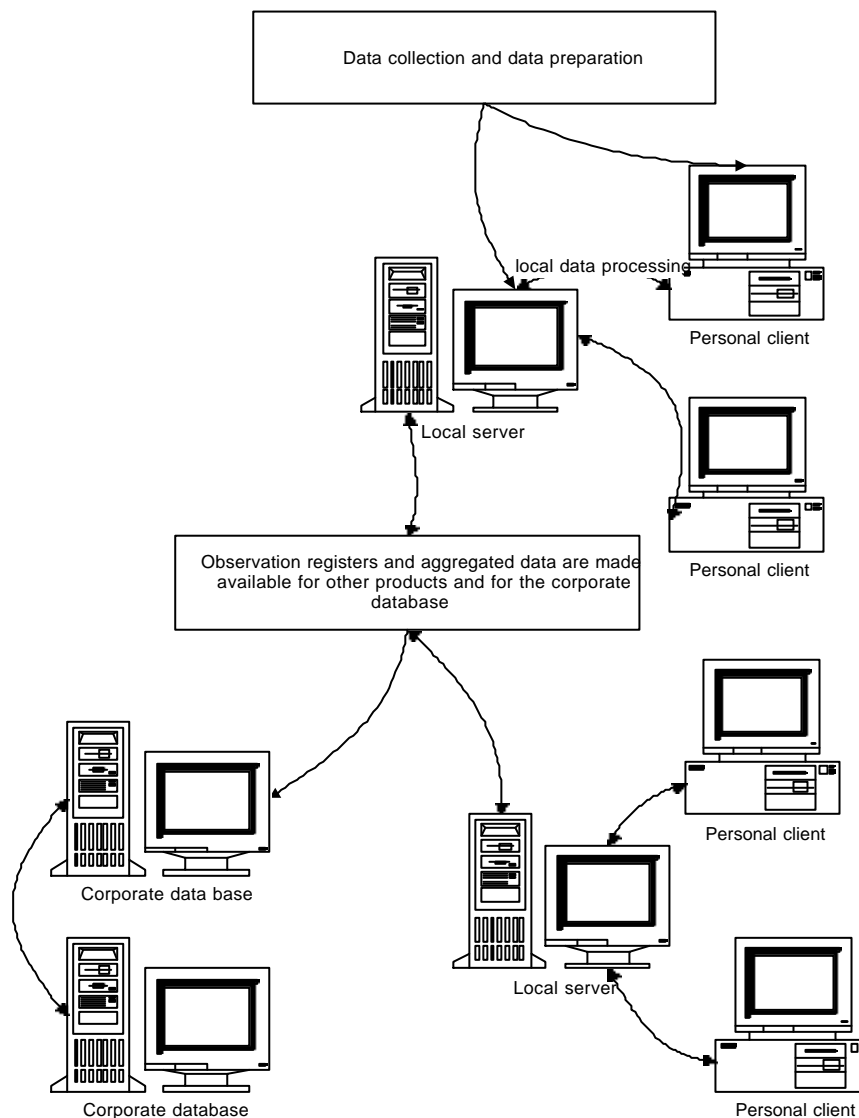
Figure 4.4 Typical distribution of tasks in a statistical information system. Data collection and data preparation tasks are done locally. Final observation registers and some predefined aggregated data are made available to the organisation as a whole.

4.3.6 Multi-tier client/server architectures and "networks of networks"

Let us analyse where to put different functions in a typical statistical information system, e.g. a survey processing system. There is typically a local working group in the statistical organisation, which is responsible for a particular survey. According to our previous analysis a client/server network supporting such a working group should consist of a client computer for each person in the working group and a server computer for some common functions for the group (and the survey) as a whole. Moreover this local client/server network should be connected to other functions inside and outside the statistical organisation of which the working group (and the survey) is a part.

How should we allocate different functions and subfunctions of the survey processing system to clients and servers? (Cf. Figure 4.5.) There are certain obvious allocations. For example, the communication interface between a single user and any computerised function handled by the user should reside on the personal client computer of the user. The same holds for truly personal applications, e.g. one-time spreadsheet calculations. On the other hand, data that need to be shared, and in particular data that need to be updated by several members of the working group during the same time period, should be stored and managed on a server computer.
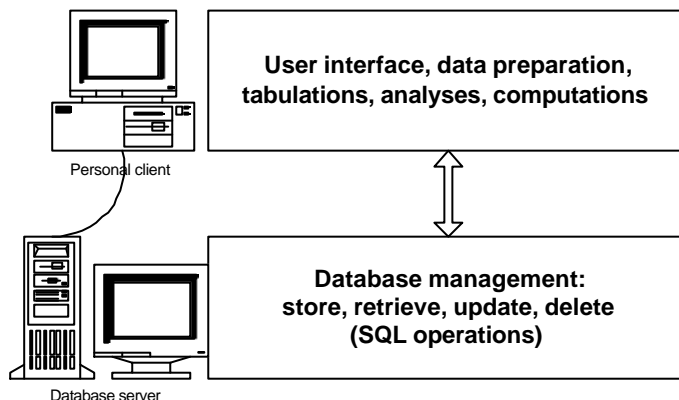
User interface, data preparation, tabulations, analyses, computations

Personal client

Database management: store, retrieve, update, delete (SQL operations)

Database server

Figure 4.5 Two-tier client/server architecture

User interface, personal applications

Personal client

Multi-user applications, "business rules"

Application server

Database management: store, retrieve, update, delete; relational operators
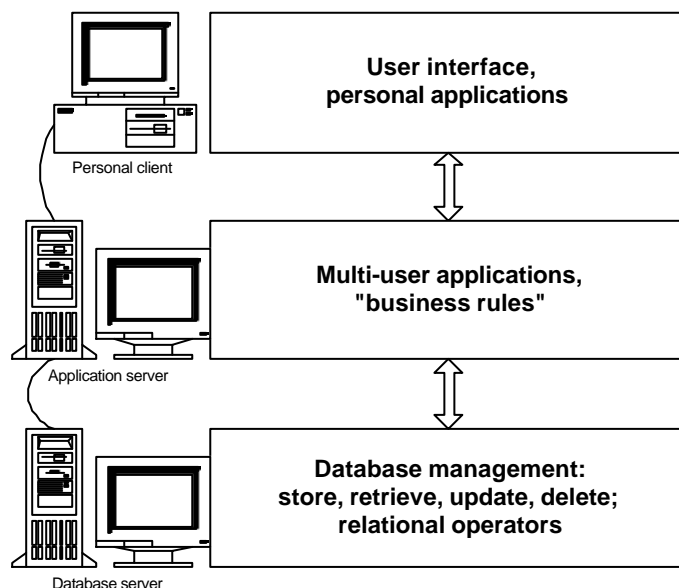
Database server

Figure 4.6 Three-tier client/server architecture

But where should we put application code, which is needed by a survey processing function that engages more than one of the group members? If we integrate this code with the common data needed by the application (e.g. by writing a so-called stored procedure) and store and manage this program/data complex on the database server, we may easily violate the principles of program/data independence that is a corner-stone of database-oriented systems design, according to what we have already discussed earlier in this report. On the other hand, if we put the application code on the client computer, we need to duplicate the code as many times as there are users of the application, and the most serious aspect of this is the maintenance problem for the code which is thus created. Moreover, if the application is a heavy one, it may block the client computer from other tasks for a considerable time period.

An attractive solution to this dilemma is to place the application code on another kind of server, an application server, and to run the application on this server on a multi-user basis. (Cf. Figure 4.7.) The application server may actually be the same physical server as the database server; the important thing is that the application and the database function are logically separate from each other and communicate through a well-defined, standard interface, like standard SQL.

The so-called three-tier client/server architecture recognises three distinct layers, or tiers, in the architecture:

- the client tier, for user interactions and personal applications;

- the application server tier, for shared applications and batch applications;

- the database server tier, for shared database management functions.

Another way of describing the distinction between the application layer and the database layer is to say that shared business rules (represented by application software) belong to the application layer, whereas shared business data belong to the database layer.

Since there may be good reasons to distinguish further layers, or tiers, within the general client/server architecture, e.g. a layer for communication with other systems, it may be more appropriate to speak about a multi-tier client/server architecture, not limiting the number of logically distinct layers to three.

When applying a multi-tier client/server scheme in a practical situation, it is important to always remember and understand why certain functions should be allocated to certain layers, and to appreciate that different logical layers may sometimes reside on the same physical server. Sometimes it may even be appropriate to split a certain logical layer between physical servers. For example, some parts of the database layer of a certain survey production system are global in the sense that they need to be accessible from other parts of the statistical organisation than the part that is responsible for the particular survey. On the other hand, other parts of the database layer of the same survey production system are local in the sense that they need only be accessible by the members of the working group responsible for the survey. In such a situation, it may be the best solution to put the local databases - which are typically temporary working databases for input data under preparation - on the same physical server as the application software. However, even if such a step is taken, it is important that the application and the data are kept logically apart, thus maintaining the principles of program/data independence. Once again, the physical integration of business rules and business data should be avoided, e.g. by making use of non-standard stored procedure code in the database management system.

In summary, Figure 4.7 illustrates a networked, client/server-based architecture for production and analysis of statistics, as seen from a technical point of view.
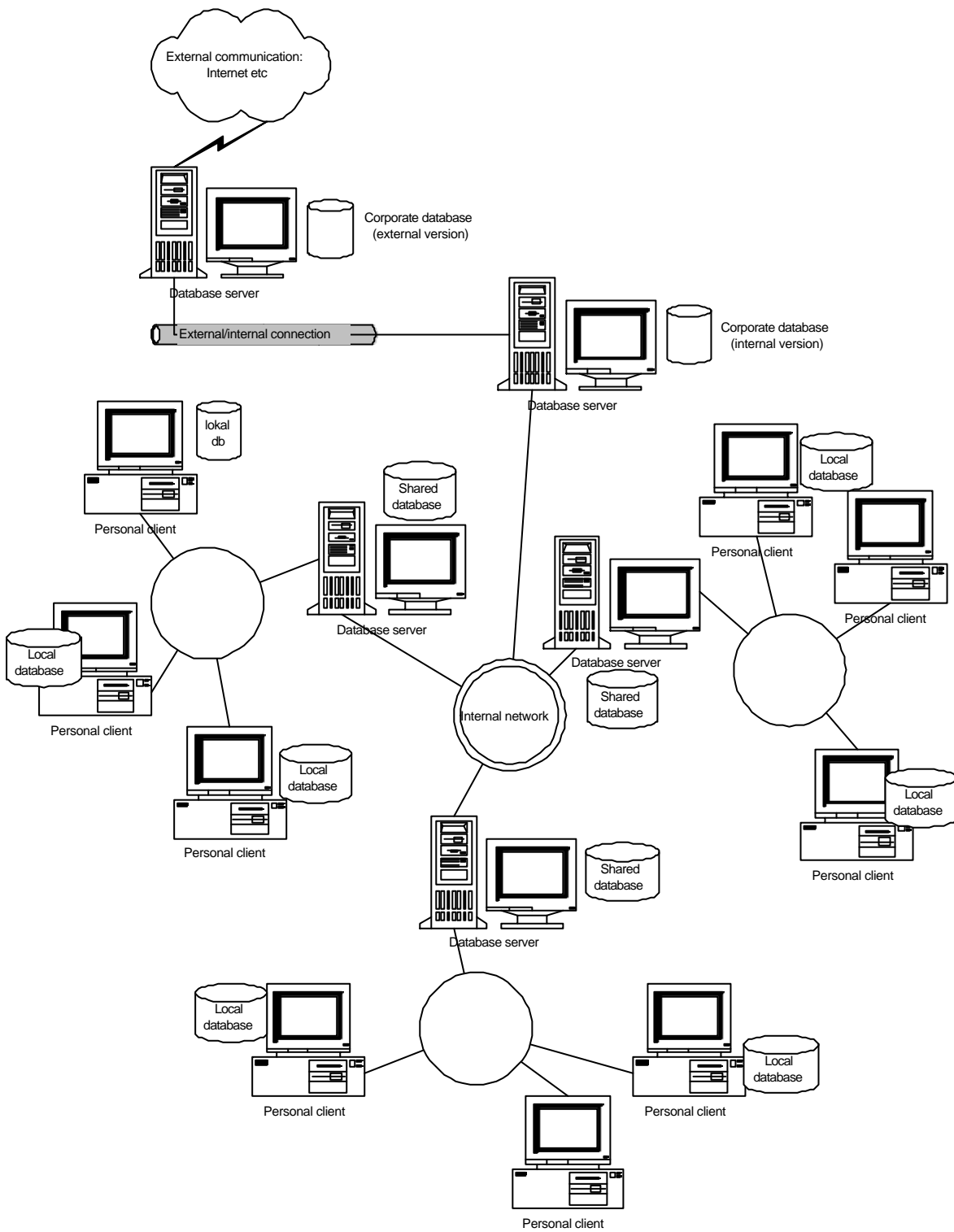
Figure 4.7 Networked, client/server-based architecture for production and analysis of statistics

CHAPTER 5

IMPLEMENTATION ASPECTS

In order to be successful in the development of an information system architecture for an organisation, one must have

- a reasonably clear vision of the target architecture to be achieved, and

- a realistic plan for how to implement the target architecture.

In this chapter we shall discuss some important aspects to be considered when planning the implementation of a new statistical information system architecture.

5.1 Aiming at a moving target

We are at the moment experiencing an unprecedented development speed in the area of information technology (IT). It is not likely that this dynamic process will slow down in the foreseeable future (3-5 years). When proposing an IT strategy and an information systems architecture for an organisation, one must seriously consider the speed of environmental changes. In particular, there must be a realistic plan for implementation, which is able to accommodate to changes that are certain to happen during the implementation process itself, although nobody is really capable of predicting the exact contents and consequences of these expected changes.

One obvious characteristic of the dynamics of IT development is that the price/performance ratios of standard hardware and software components improve all the time. Thus, it is almost always better to buy available standard components off the shelf rather than developing one's own solutions, and it is usually advisable to spend more on hardware capacity rather than complicating a simple solution.

Another implication is that standard components will come and go at a high speed. Products that were *de facto* standards five years ago may be completely swept away from the market today. So we do not really know which will be the standard components to be used in a few years from now. As a consequence, it is much safer and wiser to standardise in terms of interfaces between components, rather than in terms of the components themselves. It is true that new standard interfaces will come around, too, but it is easier and less awkward to maintain an old interface than to maintain an obsolete component, e.g. a software component. Furthermore, it is easier to replace a component conforming to a standard interface with another component conforming to the same interface than to replace a component that does not conform to any standard interface at all. It is even easier to replace one standard interface with another standard interface than to have to maintain or replace non-standard components.

One problem in very dynamic situations is how long to wait for standards or better standards. A simple illustration is the dilemma of a potential first-time PC buyer. The PC that I would like to have today will no doubt be less expensive in a couple of months, and within a year it will cost only a fraction of its price now, and there will be much more attractive PCs available at the same price as that of the PC that I am considering buying now. At any particular point in time it would seem to be better to wait a while before buying, but if you stick to this policy you will end up never buying any PC at all.

The solution to this paradox is that at any point in time you should buy state-of-the-art hardware and software components, e.g. state-of-the-art PCs, but you should write them off as quickly as possible. More importantly, you should buy components that conform to market

standards, so that you can replace the component with a better one, without having to change any other components (hardware or software). Even if the market standards should change, there will certainly be a market for providing upward compatibility from old standards to new ones.

In addition to having defined a number of strategically important interfaces, it is advisable for an organisation to have a clear picture of its overall information systems architecture. A standard information systems architecture, based on standard interfaces, standard subsystems, and standard component types (with well-defined functionalities), and (of less significance) standard components, will provide a certain amount of stability and flexibility to the organisation, even in times of extreme IT volatility, when you always have to aim at a moving target.

Let us discuss these principles in more concrete terms. As regards hardware, the Intel processor and the IBM-compatible PC are well-established standards today. These standards will continue to be improved, but they are not likely to become obsolete during the next 3-5 years.

On the software scene, Microsoft is the undisputed setter of standards and trends.[1] Thanks to its financial strength, if nothing else, Microsoft will be a safe bet for the next 3-5 years, even if miracles (positive and negative) have happened before in the IT world. In some

ways Microsoft now occupies a position in the PC world similar to that of IBM when mainframes prevailed. However, there is an important difference between IBM in those days and Microsoft today. IBM always wanted to lead the software development, and it systematically tried to lock out competitors and lock in customers by making their systems very closed. Microsoft has taken another approach. Although they favour their own standards over negotiated standards, they actively support competitors who want to develop software products compatible with Microsoft standards. This can be seen from the fact that other software developers have often been the first to launch very innovative Microsoft-compatible products, and these pioneering companies, e.g. Lotus, have usually had some commercial success, before Microsoft comes out with a similar product. Microsoft always aims at the mass market. This is reflected in the pricing policy, which is actually very much to the benefit of even big software-buying organisations.

It should be stressed that, today, it is often more important to evaluate the commercial strength of a software producer than to evaluate technical details of the software products as such. If a software producer is commercially stronger than its competitors, it will soon catch up with any technical advantages that its competitors may have anyhow.

5.2 Short time horizon and realistic ambitions

Due to the rapid pace of general progress in the IT area, there is no point for an organisation to start development projects lasting several years. During the lifetime of, say, a four-year project, one is likely to find after, say, two years that the system under development is rapidly becoming obsolete. In such a situation, one has the choice between two bad alternatives. One alternative is to complete the development, as planned, which (at best) will lead to the result that the organisation two years later has a brand new, but already old-fashioned, system. The other alternative is to start the project over again, sacrificing two years development work;

[1] A warning is appropriate here. This paragraph may sound like an advertisement for Microsoft. However, it seems to be the rule rather than the exception that the IT market is (at every particular point in time) dominated by strong market leaders, who set de facto standards. Today a statistical agency is a very small player on the IT market and can usually do nothing better than follow the main stream. This strategy, if followed by many, may seem to be risky in the long run, since it may tend to perpetuate and strengthen unhealthy monopolies. However, no monopoly in the IT market seems to survive for ever. Even IBM was about to go bankrupt a few years ago. No doubt the Microsoft monopoly will also have its time limit, and somebody else will take its place. However, whoever is the market leader for the time being, the principles suggested here will remain valid for statistical agencies to follow.

hopefully the planners will choose a much shorter time frame for the new project.

Even large and complex projects should not be permitted to have a time frame of more than one or maximum two years. Such projects should be divided into subprojects with clearly defined results, milestones and very sharp deadlines. If a deadline is threatened, it is almost always better to modify ambitions than to postpone the deadline. This strategy requires project managers to be alert, and to always consider which are the top priorities right now; this is useful, since priorities tend to change over time in a dynamic environment. Neither should project managers be afraid to give up some decisions now and then. Mistakes are inevitable, and new and better tools come around all the time. It would be stupid not to exploit new tools and products, if they are significantly better than the old ones, even if it sometimes leads to some "sunk costs".

This brings us to the important question of whether one should aim at "quick and dirty" rather than "perfect" solutions. The answer is that neither of these alternatives is satisfactory. There is no use in aiming at "perfect" solutions, because they will take too long time to implement, and when they are ready, they are no longer "perfect" in the new situation. "Quick and dirty" solutions are no good either. It is true that too many over-ambitious goals should be avoided, but those goals which end up with high priorities after careful considerations and reconsiderations should be based on sound principles, so that other goals can be added incrementally by subsequent projects.

When a statistical organisation plans for migration to a new technical platform, or when it organises a data warehouse, based upon modern techniques like databases and the Internet, it is not unusual for staff members to suggest taking the opportunity to improve the contents and quality of statistics at the same time. To some extent this may be done without too much extra effort, but as soon as such activities threaten the time schedule of the project, the critical question of how important these improvements really are, and which

priority should they have, must be asked. After all, statistical organisations have usually survived for a long time, and even with a high public reputation for good quality, although deficiencies have existed, e.g. as regards the co-ordination of data from different surveys. New and better technical solutions will no doubt increasingly expose these deficiencies, and this may be acceptable. In the long run such feedback from users may be of utmost importance for sustainable improvements of the quality of statistics, and sustainable improvements are more valuable than temporary patchwork.

5.3   Organisation and control

The statistical information systems architecture proposed in this report is not dependent on the organisation as such. In principle, the information systems architecture could be the same regardless of, say, which degree of centralisation/decentralisation is chosen for the organisation of statistical responsibilities. Similarly, the proposed statistical information systems architecture could co-exist with different organisational solutions for IT-related work.

There are two major alternatives for how to carry out a transition to a new statistical information systems architecture. One major alternative is a centralised one. Top management would organise a special project, with special resources, for planning and implementing proposed changes all over the organisation. The other major alternative is a decentralised one. Top management would have to be engaged in this approach, too, and a special project would have to be organised. However, this project would not have any responsibility for implementing individual applications; this responsibility would rest with the departments that are responsible for the respective applications. The central project would have the responsibility for setting standards and for supporting application development in various ways, for example by solving certain problems of a general nature, like the development of certain generalised software tools, as well as acquisition of

recommended commercial software tools and components.

In many statistical organisations, the second alternative is the only realistic one. The first alternative requires a lot of central financing and co-ordination, and it runs obvious risks of being delayed for various reasons. The second alternative may not lead to complete success in all parts of the organisation - but neither does it run the same risks of a complete failure as the first one. Those departments that want to move fast could do so, without having to wait for latecomers. Nevertheless, it is of course important for the organisation as a whole that the migration process is finalised everywhere as soon as possible. It is expensive to be in a transition stage longer than necessary.

Top management engagement in this type of project is essential. On the other hand, top management needs support from the organisation. A staff function, assisted by the central IT department, could provide such support. The project should focus on statistical information systems and statistical tasks, but IT is of course a major instrument that the project has at its disposal. Possibilities to improve statistical co-ordination should be noticed and actively exploited, e.g. by means of the global metadata component of the data warehouse.

Production and analysis of statistics is the core business of a statistical organisation. Thus, it should be given priority when IT resources are allocated. Like other organisations, statistical organisations use IT for administrative processes as well. Administrative processes should serve and support core business processes. It is the needs of core business processes that should mainly influence the design of the information systems architecture of a statistical organisation, including the technical infrastructure. In order to stimulate this, it is advisable that the central IT function be financed indirectly through its customers. The different departments of a statistical organisation should be given money as part of their own budget for buying services from the IT department. This procedure is particularly important for application-oriented work. For such work the departments could be given a free choice as to whether to have their own IT staff or whether to buy services from the IT department. Other services, e.g. printing, could be controlled by its customers in a similar way, i.e. there should be an internal cost recovery system. This is important in order to overcome suboptimisations that may otherwise occur.

The central IT function has an important role in the development of generalised IT tools and solutions. As far as the development of the statistical information system infrastructure is concerned, the resources required for this work could be financially controlled by a strategically oriented staff function, on behalf of top management.