MAKING STATISTICAL DATA MORE AVAILABLE

Bo Sundgren

SCB

R&D Report
Statistics Sweden
Research - Methods - Development
1995:6

INLEDNING

TILL

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

Föregångare:

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

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

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

Efterföljare:

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

R & D Report 1995:6. Making statistical data more available / Bo Sundgren.
Digitaliserad av Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-1995-X101OP9506

MAKING STATISTICAL DATA MORE AVAILABLE

Bo Sundgren

SCB

R&D Report
Statistics Sweden
Research - Methods - Development
1995:6

Från trycket December 1995
Producent Statistiska centralbyrån, utvecklingsavdelningen
Ansvarig utgivare Lars Lyberg

Förfrågningar Bo Sundgren
tel 08-783 41 48
telefax 08-783 45 99

ISSN 0283-8680

Printed December, 1995
Producer Statistics Sweden
Publisher Lars Lyberg

Inquiries Bo Sundgren
telephone +46 8 783 41 48
telefax +46 8 783 45 99

ISSN 0283-8680

MAKING STATISTICAL DATA MORE AVAILABLE

Bo Sundgren
Statistics Sweden
S-115 81 STOCKHOLM
Sweden

Summary

Will statistical offices be able to meet new challenges from the users to make statistical data more available by means of modern technology? Can they to do this within existing budget restrictions, and with due consideration to the interests of data providers? These are questions addressed here. Problems and opportunities are illustrated by examples from Sweden.

1 New challenges for statistics producers

Statistics producers in national statistical offices are facing new expectations, demands, and requirements from several directions:

- from *statistics users*, who want faster, easier, and less expensive access to statistical data - through media and routines that are better adapted to their own processing needs;
- from *data providers*, who demand less burdensome reporting - through media and routines that are better adapted to their own information systems;
- from *governments and tax-payers*, who want "more value for less money";
- from *international organisations*, requesting member countries to provide timely, comparable, good quality statistics, which comply with international standards.

Technological progress is taking place as rapidly as ever. All the above-mentioned stake-holders in statistics production expect statistics producers to take full advantage of advances in technology. This paper will discuss how statistics producers can respond to some of the challenges. The paper focuses on how statistical offices can make statistical data more available to statistics users, while satisfying restrictions given by scarce resources and the willingness of data providers to co-operate.

2 User-orientation and user-friendliness

There is a need to review the concepts of user-orientation and user-friendliness. It has become a widely accepted dogma that information should be user-oriented and user-friendly. All information system designers pay lip services to this dogma. To be fair, most designers sincerely believe they are developing systems characterised by user-orientation and user-friendliness, although they have since long stopped thinking more deeply about the meaning of these concepts.

In the early ages of computer usage, that is in the 1960's, the direct user of a computer had to be a computer programmer. Since most computer applications in those days were mathematically

oriented (as suggested by the word *computer* itself), it meant a step forward from the user's point of view, when the user/mathematician could communicate with the computer by means of mathematical formulae (like in FORTRAN) rather than having to program in machine code or assembler languages. The programming language COBOL meant a similar step forward for users/programmers oriented towards administrative applications.

In a statistical office there are numerous information systems applications of more or less the kind: statistics production. As systematised by figure 1, a statistical production process includes a number of very typical functions like frame administration, sampling, data collection, data entry, coding, editing, estimation, tabulation, analysis, and presentation. In the late 1960's there were few other organisations, if any, which had a similar opportunity to exploit economies of scale in the development of computer applications. Thus, not surprisingly, statistical offices became pioneers in the development of generalised software. These software products often supported high-level, non-procedural command languages, which enabled non-programmers to develop applications within a certain application area by simply specifying

(i) the input data to the application, e.g. a so-called flat file with a certain record layout; and

(ii) the requested output from the application, e.g. a statistical table with a certain contents and a certain layout.

The variability of applications developed with tools of this type has to be relatively limited. This condition is satisfied by the functions corresponding to production steps of a typical statistical survey.

The high-level, non-procedural command languages represented a certain degree of end-user orientation in a computing environment that was based upon mainframe computer centres operated as closed shops and in batch mode. In the early 1970's user-orientation and user-friendliness became more or less synonymous with person/computer interaction through menu-driven information systems. Certainly these systems helped to bridge the gap between the computer and its non-programmer end-users. Nevertheless it was still very much the computer that controlled the user rather than the other way around. The user could choose his route through the hierarchy implied by the menus of the menu-driven system, but he could not affect the hierarchy as such, and he had to go through the hierarchy level by level in a rather rigid way.

The introduction of powerful, inexpensive micro-computers in the beginning of the 1980's added several new dimensions to the concepts of user-orientation and user-friendliness. First of all the new technology meant that the closed mainframe shops could be closed for good as far as many of the users were concerned. The users suddenly found themselves in control of computer resources in much the same way as they already were in control of other resources necessary for their daily work. The computer became demystified. Furthermore, the new technology finally enabled the user to take control of the computer rather than the other way around. This possibility materialised in the windowing techniques pioneered by Xerox, followed up by Apple, and successfully mass-marketed by Microsoft.

Today practically every user of statistics is a user of computers as well. He has his own computer in the office, at home, and when travelling. He demands to choose whatever software he prefers to retrieve, process, and analyse statistical data. Through standardised network services (in his own office as well as world-wide) he is able to communicate and co-operate with other human beings and other computers, and he is able to do this very much on his own conditions.

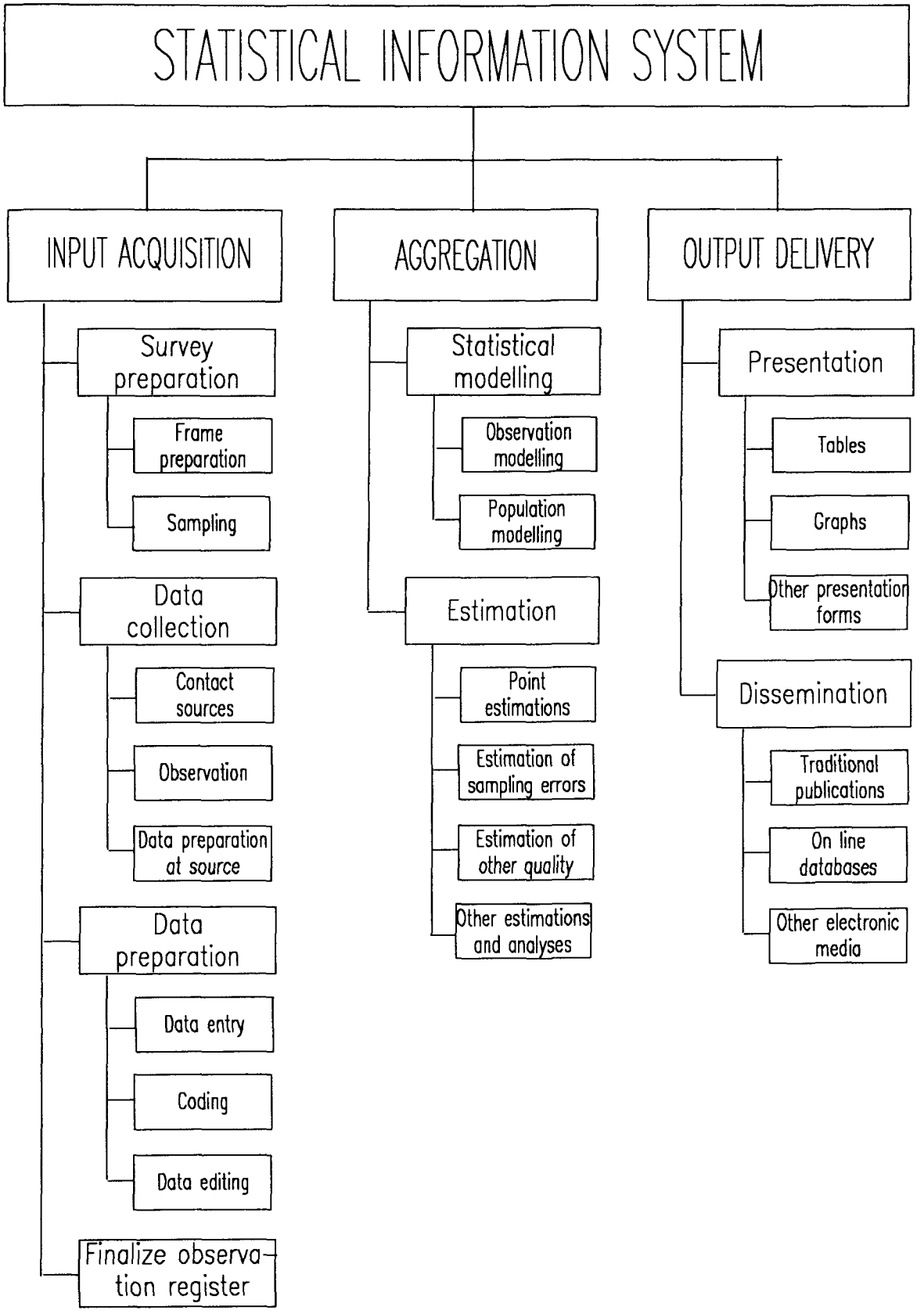

Figure 1. *A functionally oriented model of a statistical information system.*

Naturally, in this situation there is not - and cannot be - a single concept of user-orientation and user-friendliness. Different users have different needs, different resources, and different preferences. There are indeed a wide variety of user profiles, as suggested by figure 2. It would be futile for a statistical office to try and satisfy all these different requirements with one and the same notion of user-orientation and user-friendliness. On the other hand, it would be equally futile to try and tailor specific products and services for each potential user of statistics. The challenge for a modern statistical office is to offer a multitude of products and services ranging from

- simple free-of-charge products based on self-service; over
- standard, off-the-shelf product/service packages charged according to price-lists; to
- sophisticated, tailor-made services provided to individual customers on the basis of tenders.

3 Standard interfaces: decreased complexity and increased flexibility

It is a challenge for a modern statistical office to be responsive to expectations, demands, and requirements from an ever more dynamic environment. Society itself, which is to be reflected by statistical data, is changing at an ever faster rate. This leads to needs for more variability, more flexibility, on the input side as well as on the output side of statistical information systems managed by statistical offices.

In order to manage requirements for greater variability in the exchange of data with the external world, and in order to do this with the same or even less financial resources, a statistical office must consider system level actions. It is not enough just to do "more of the same thing" or to "run faster". It is necessary to undertake more drastic redesign actions.

Making more extensive and more systematic use of standard interfaces are actions that may lead to desirable system changes. Such actions may lead to a combination of the following two consequences:

- a drastic decrease in the complexity of data exchange between statistical information systems and their environments as well as between the internal components of the individual statistical information systems themselves;

- a drastic increase in the (actual or potential) variability and flexibility in the (external and internal) behaviour of the statistical information systems.

Both types of consequences are highly desirable. Figure 3 from Malmborg & Sundgren (1994) illustrates the differences in terms of complexity and variability between

- a situation where two sets of systems interact directly in the absence of a standard interface (figure 3a); and

- a situation where the same two sets of systems interact via a standard interface (figure 3b).

USER CATEGORY BY CHARACTERISTIC	"Ministry of finance"	"Researcher /scientist"	"Analyst - public sector"	"Analyst - private sector"	"Actor on the finance market"	"International organisation"	"Journalist"	"Politician"	"Teacher/ student in school"	"Interested citizen"
Competence: - subject matter - statistical - EDP										
Knowledge about relevant data sources: - broad - deep										
Quality requirements: - contents - accuracy - availability										
Needs for search systems, documentation, and metainformation										
Resources: - hardware - software - expertise - money - "trading objects"										

Figure 2. *A scheme for analysing the profiles of different categories of statistics users.*

In the situation illustrated by figure 3a, the interaction format will have to be negotiated for each combination of systems that need to interact. This will typically lead to many different, tailor-made interaction formats that require a lot of resources to develop and maintain. The situation is inconvenient from operation point of view as well, since every individual actor will have to remember different interaction formats for different interaction partners. If a new system is added to any of the two sets of systems, a new interaction format will have to be negotiated for each other system, with which the new system needs to interact.

In the situation illustrated by figure 3b, every system will need to develop, maintain, and operate one single interaction process, the interaction with the standard interface. Through this process, every system will be able to communicate with all other systems, including systems that do not yet exist but will be introduced later. Thus, in comparison with the situation in figure 3a, this situation is both less complex (to develop, maintain, and operate) and more flexible vis-à-vis growth and other changes in the system environment.

Figure 4 indicates a number of places where a statistical information system could and should contain well designed, preferably standardised interfaces. One may distinguish between

- external, inter-system interfaces; and
- internal, intra-system interfaces.

External interfaces are interfaces between, on the one hand, the statistical information system under consideration and, on the other hand

- statistics users: human end-users as well as other (statistical) information systems; these are output-oriented interfaces;

- data providers: human respondents as well as other (administrative) information systems; these are input-oriented interfaces.

An example of an output-oriented standard interface for statistical information systems is the GESMES format for representation of statistical macroinformation and accompanying meta-information. "GESMES" stands for "GEneric Statistical MESsage", and the standard is developed by the UN/EDIFACT Message Development Group 6.1.

Similarly, on the input side, there are several UN/EDIFACT standard formats corresponding to typical documents of different branches of activity in society, e.g. trade. A generic standard for input messages to statistical information systems is the Raw Data Reporting Message; see UN/EDIFACT (1994).

By providing a statistical information system with standardised external interfaces, the designer makes the system open and easy to integrate with other systems, e.g. the local systems of users and providers of statistical data. This is indeed a practical application of the theoretical principles illustrated in figure 3 above. By accepting data and metadata through standardised interfaces, a statistics producer facilitates for respondents to provide statistical raw data as a natural side effect of their own administrative routines. Analogously, by making (aggregated or anonymised) data and metadata available through standardised interfaces, a statistics producer facilitates for statistics users to integrate statistical data from the statistics producer with the user's own (statistical and administrative) data for analyses and decision-making.

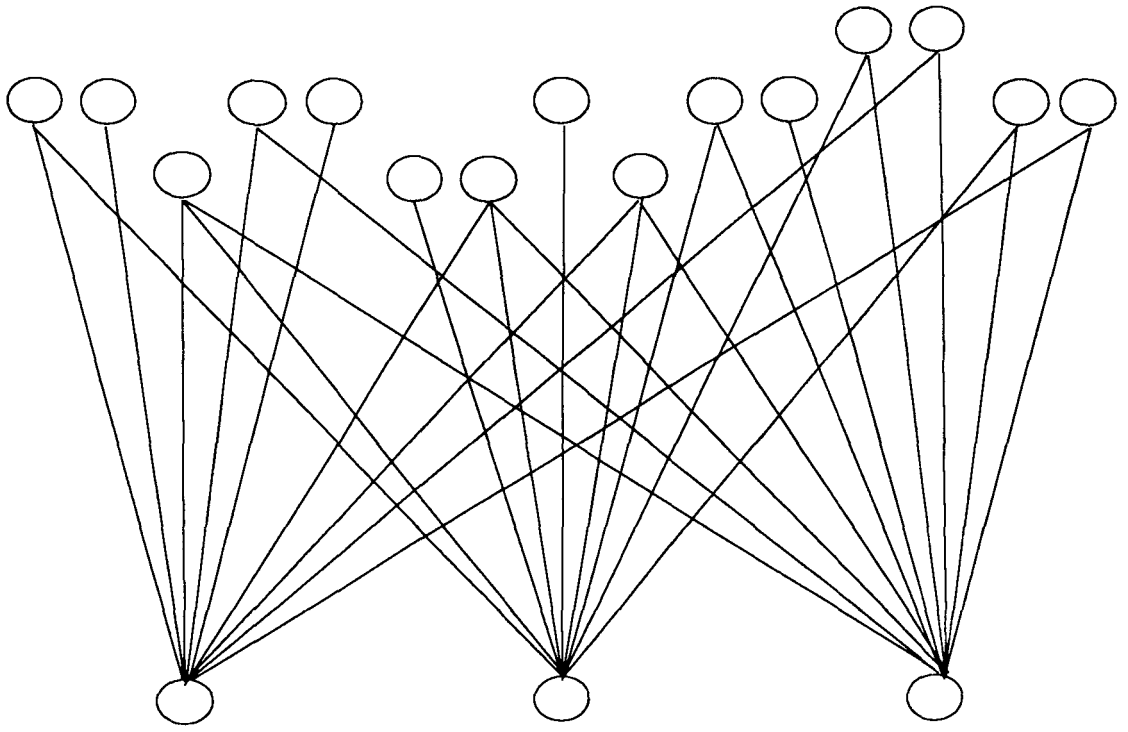

Figure 3a. *One way of organising the interaction between two sets of systems.*

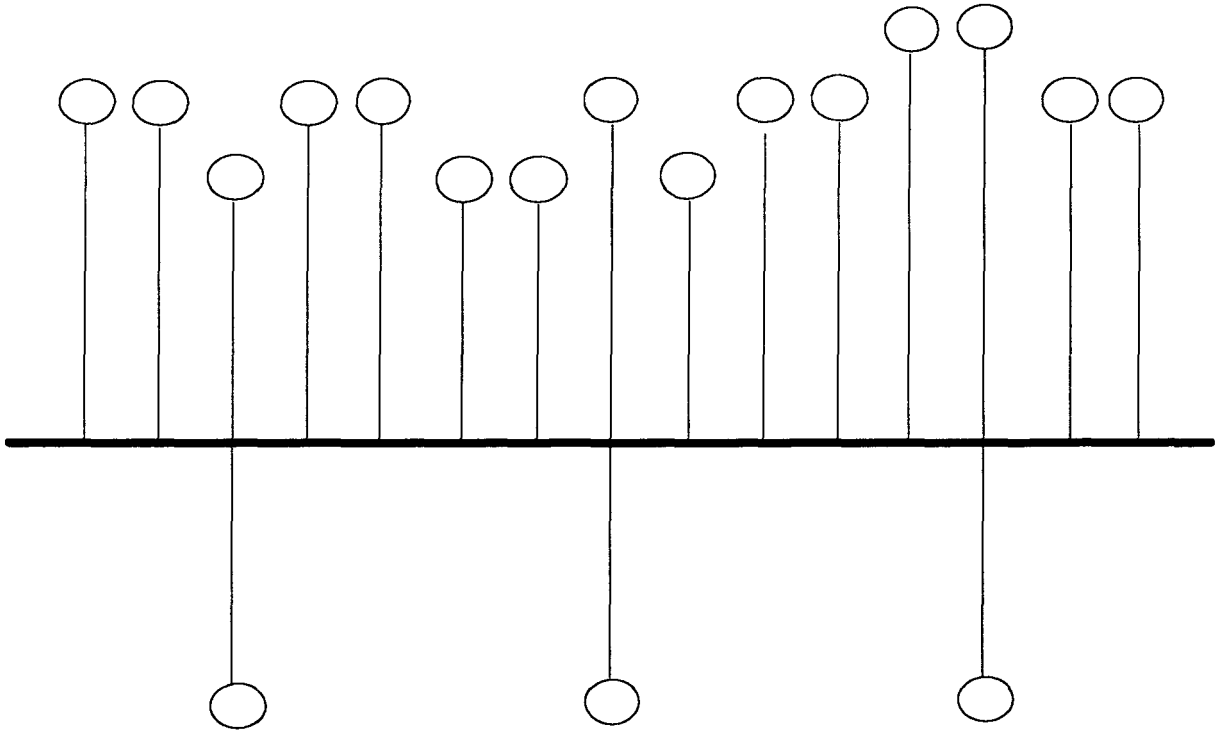

Figure 3b. *Interaction between two sets of systems via a standardised interface.*

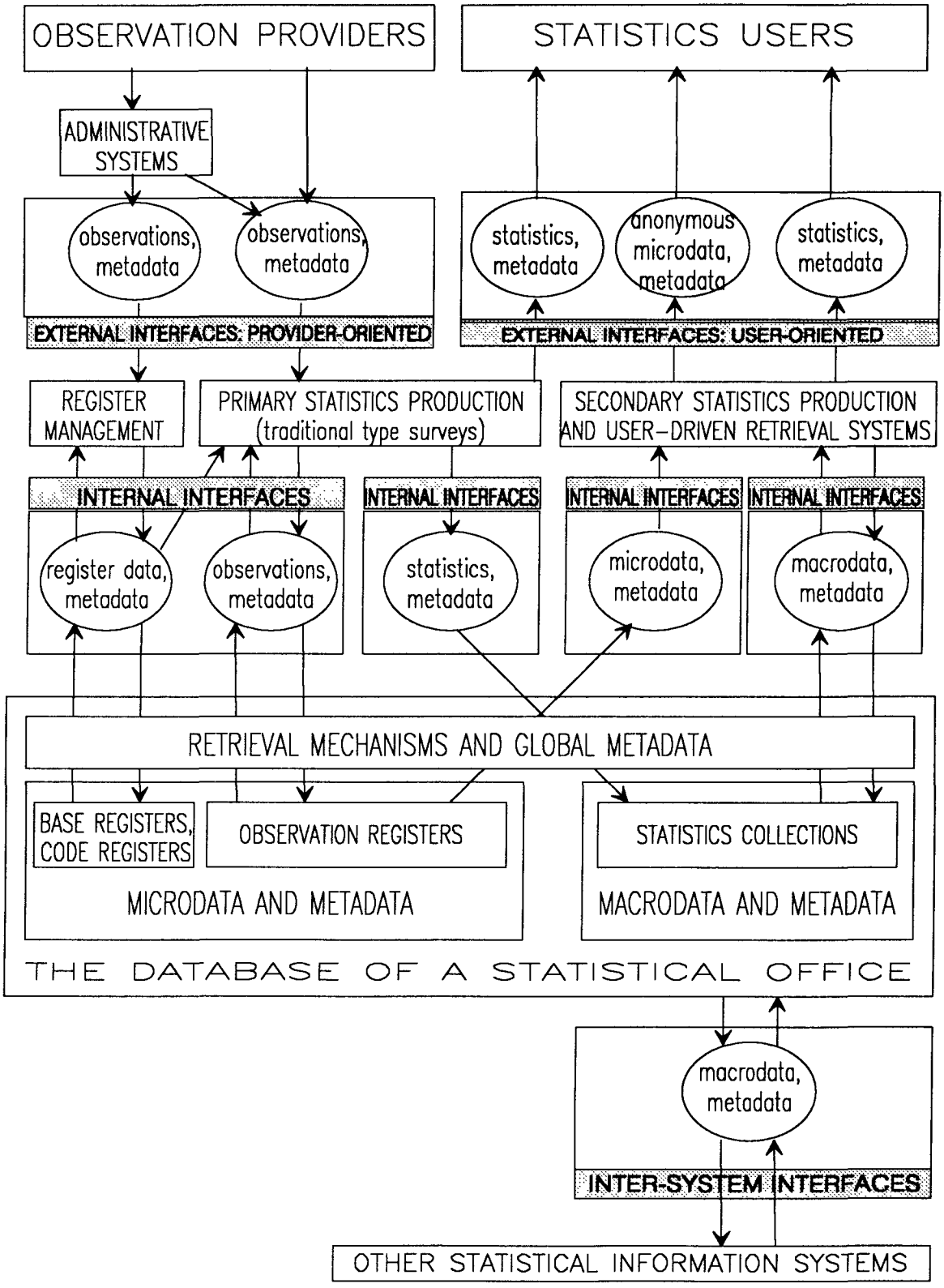

Figure 4. *A database-oriented statistical information system with clearly defined internal and external interfaces.*

Statistical offices began to realise the importance of standardised internal interfaces, at least implicitly, when they started to exploit the benefits of generalised software at a large scale in the middle of the 1970's. As long as statistical information systems were completely tailor-made by professional programmers, who were using procedural programming languages, there was not a strong enough incentive to define and use standardised interfaces between software components. It was up to the individual programmer to define suitable data structures as well as formats and procedures for data interchange. When generalised software products gained in popularity, much on the initiative of non-programmers, one problem was the enormous variability in data structures and data interchange formats and procedures that were exhibited by existing applications and data files. It was first considered to further develop the generalised software tools in order to make them capable of handling this variability. It was soon realised that this would be a Sisyphus task. Instead some statistical offices decided to standardise data structures on the basis of the concept of a "flat file", that is, a file containing only one record type, adhering to a record layout with a fixed number of fields containing the (single) values of the attributes, or variables, of one particular instance of a certain object type, e.g. a person, a household, or an enterprise. Multiple record types, hierarchical records, and repeating groups were among the data structure phenomena that were banned in this standardisation process.

This standardisation of data structures and data interchange can be seen as a first step towards database-oriented information systems. Technically speaking, there was no physical database visible in those systems, where data were stored and exchanged in sequential files stored on magnetic tapes. Nevertheless the "flat file" standard started to play the same role as the relational data model (with the SQL interface) has in today's database-oriented systems. Different processes, controlled by different generalised or tailor-made software products, exchanged data as flat files - within and between statistical information systems. The generalised software products were often developed within the statistical offices themselves, but the same principles could easily be applied to commercial software as well. In fact commercial software could very seldom handle more complex data structures than flat files anyhow.

In a modern statistical information system the relational data model and the SQL standard for data interchange between application software and the database management system are obvious choices for internal interfaces. All commercial software products that want to survive on the market have to adhere to these standards. Another *de facto* standard (though limited to PC software) is Microsoft's Object Linking and Embedding (OLE) for transferring data and control between different software components.

Figure 5 indicates how the different functions of a statistical information system (cf figure 1) could be designed to interface the database including microdata, macrodata, and metadata.

No standards are for ever. Maybe in five or ten years time today's *de facto* standards will have become replaced by others, e.g. a widely accepted standard for object-oriented database management. This is not a great problem. It is relatively simple to move from one standard to another. It is much more difficult to live in a non-standardised situation, and to make the first-time move to a standard. Nor does it matter very much if standards are formally agreed upon by standardisation bodies. What is critical is that standards should neither discriminate software manufacturers from taking part in competition, nor force software users to be faithful to any particular hardware or software vendor.

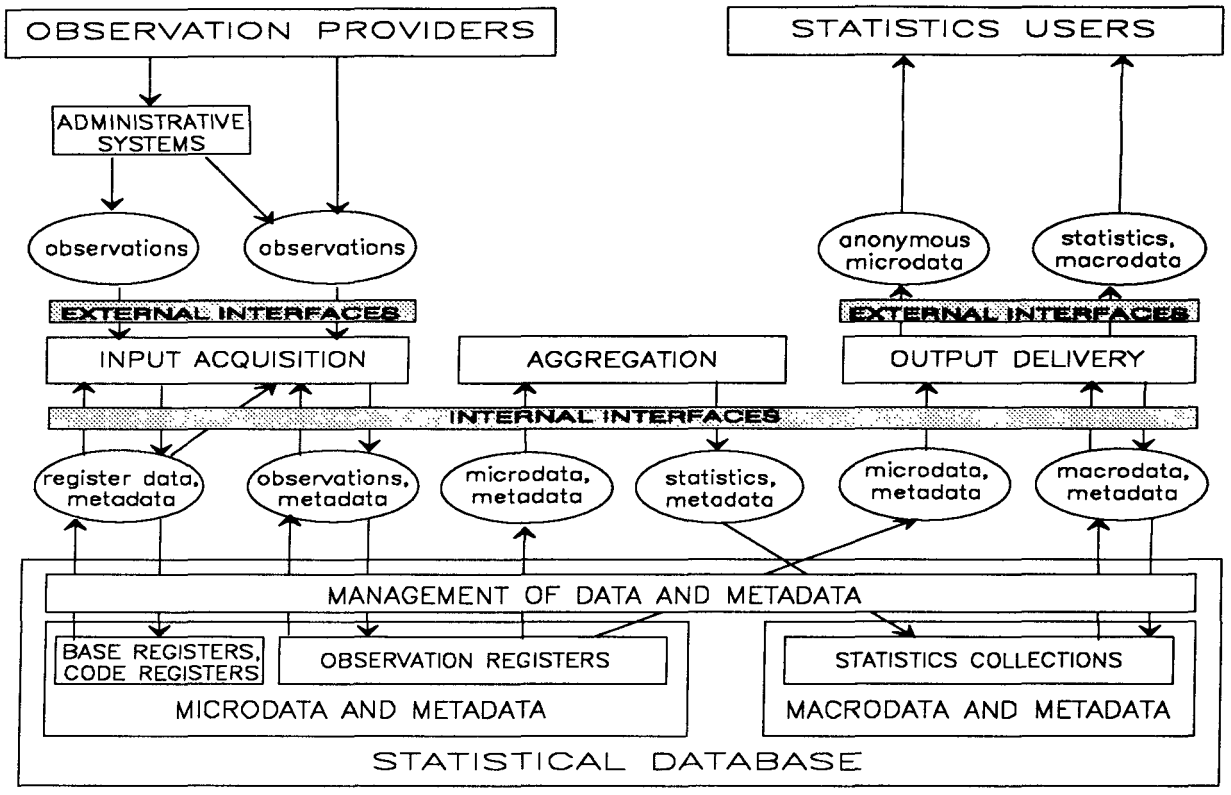

Figure 5. *A functionally oriented model of a database-oriented statistical information system.*

4 Standard components: off-the-shelf software

Statistical offices were among the first companies and organisations to make systematic use of standard components (e.g. generalised software) in the development of information system applications. Already during the sixties statistical offices started to use commercially available and/or in-house developed statistical packages for common statistical operations like data editing, tabulation, and statistical analysis. During the seventies some statistical offices could start reducing the number of application programmers, encouraging subject matter statisticians to develop (part of) their own applications by means of high-level, non-procedural, generalised software tools. This development was intensified during the eighties.

With the advent of inexpensive PC technology and software, the boundary between "user programming" and "professional programming" has become blurred - in statistical offices as well as in the data processing community at large. Major companies are closing down their central application development departments, advising business departments to use ready-made software packages for auxiliary functions, and to "put together" business-critical applications from software components that can be bought off-the-shelf from commercial software vendors.

Welke (1994) has predicted that we shall see a paradigm shift in how information systems are typically developed:

"There is a fundamental paradigm shift underway in how (information) systems and the software which supports them, is developed. The shift is away from a craft-based structure in which user requirements are specified and custom solutions developed, to a market-product based approach in which the users themselves select and arrange meaningful-to-them components as a solution to their requirements."

The paradigm shift is likely to imply an even greater future for such things as

- inexpensive, generalised software, availble "off-the-shelf"
- "tool-boxes" containing generalised standard components
- rapid application development (RAD) methods and tools

In connection with RAD, it should be noted that tools for Computer-Assisted Systems Engineering (CASE) are likely to become more domain-specific than today. Jackson (1994) has articulated the importance of domain-specific knowledge for software development:

"The large aspiration to place the whole of software development ... as one more branch of engineering is misconceived. Our aspiration should be to develop specialised branches of software engineering ..."

" ... there are no casual builders of cars or bridges- But in software development it is not easy to draw a clear line between the casual developer and the serious, professional developer. As a result, ... software development is still largely an amateur activity in a very important sense."

5 Metadata

There are many potential users of statistical data in a modern society. Many of them have the competence as well as the hardware and software resources needed to take full responsibility for their own usage of statistical data. They are eager, and sometimes impatient, to exploit the information potential of statistical offices, and to do this on their own conditions - as far as permitted by confidentiality restrictions. One major obstacle, which often prevents them from doing so, is the present inadequacy of available metadata, that is, the absence or inadequacy of systematic descriptions of statistical data and the processes behind them.

A (potential) user of statistical data will need metadata for three major purposes:

1. searching for potentially relevant and useful statistical data;
2. evaluating the adequacy of available data and the cost/benefit of using them;
3. retrieving, interpreting, and analysing statistical data.

First, statistical metadata are needed as a basis for search operations. The (potential) user is looking for statistical data that could be relevant and useful for him in describing, analysing, or solving a certain problem. The traditional approach is for the user to turn to a statistical office. Staff members of statistical offices are often very helpful, but today this approach is not sufficient. There are far too many potential users for any statistical office to cope with face-to-face. In addition, many users need to combine statistical data (and other data) from several sources, and no particular staff member, or even organisational unit, of a statistical office will have the necessary overview. Moreover, manual help-functions are relatively expensive and slow, even if

they are computer-assisted. Today a user will expect the metadata needed for search tasks to be organised and disseminated in such ways that he himself can search for relevant data on the basis of widely available, computerised metadata. The process may start from a relatively vaguely expressed information need. The computerised, metadata-supported process should help the user to better understand his own needs, and it should result in explicit references to available statistical data, which are likely to be relevant for the user's problem.

Second, once the user has identified some statistical data of potential relevance for his problem, he will have to determine, if the data are really adequate for the intended purpose. This means that the user has to evaluate the quality of the data, and to consider whether it is really worth the effort and cost to retrieve, interpret, and analyse the data.

Third, if and when the user has come to the conclusion that certain available data are of sufficient quality to justify the efforts and costs to use them, he will need metadata in order to actually retrieve, interpret, and analyse the data. Retrieval may be accomplished by downloading data and accompanying metadata to the user's own PC or by obtaining a disk or CD-ROM copy. Interpretation and analysis will require the same kind of metadata as were needed for making the preliminary judgement of the quality of the data. However, at this stage it may be necessary to obtain deeper and more precise information about how the data were collected and processed, before they resulted in the available statistics.

The documentation templet in figure 6 identifies metadata items that are desirable or even necessary as a basis for responsible usage of statistical data emanating from a particular statistical survey. If appropriately compiled with the corresponding metadata for other surveys they may also serve as a basis for search operations. The survey documentation templet is part of the documentation system SCBDOK, developed by Statistics Sweden. See also Sundgren (1991a, 1991b, 1992, 1993a, 1993b).

It is an equally important task for a statistical office to produce metadata concerning its surveys as to produce the survey data themselves. In order to be able to accomplish this task in an efficient way, the statistical office must carefully design its metadata flows. Metadata should be captured when they naturally arise for the first time, e.g. as the result of a design decision. At later stages it should be possible to have them automatically transferred and transformed when survey data are transferred or transformed. Furthermore, it should be possible to have the metadata consistently updated, when the survey processes are changed, e.g. as the result of new design decisions.

The metadata describing a statistical survey and its data outputs are a combination of formalised metadata, e.g. code lists and record descriptions, and free-text metadata like verbal descriptions of variables and processes. Thus software systems for handling statistical metadata may require different types of software components to be combined, e.g. relational database management systems and software for managing and searching large amounts of text data. Hypertext software (like in advanced help functions and high-level Internet-tools) will also have a great potential for enabling the users to navigate and associate in available statistical data and metadata and to process them in efficient and intelligent ways.

DOCUMENTATION TEMPLET FOR A STATISTICAL SURVEY

0 Administrative information

0.0 Documentation templet
0.1 Survey name and identification, organisation and persons responsible
0.2 Documentation modules and subsystems
0.3 Archived data sets and published statistics
0.4 References to other relevant documentation

1 Survey contents

1.1 Domain of interest and target domain, verbal description
1.2 Target domain, formal description
- 1.2.1 Target objects, description and object graph
- 1.2.2 Target populations
- 1.2.3 Target variables

1.3 Survey outputs
- 1.3.1 Structured overview of the tabulation plan
- 1.3.2 Publications in printed form
- 1.3.3 Electronic distribution
- 1.3.4 Database storage

2 Survey plan

2.1 Frame procedure and observation objects
- 2.1.1 Overview
- 2.1.2 Frame and its links to objects
- 2.1.3 Frame production
- 2.1.4 Overcoverage and undercoverage

2.2 Sampling procedure (if applicable)
2.3 Data collection procedure
- 2.3.1 Observation objects, description and object graph
- 2.3.2 Data sources, including contact procedures
- 2.3.3 Observation variables and measurement instruments
- 2.3.4 Interruptions (including actions at overcoverage)
- 2.3.5 Non-response actions

2.4 Planned data preparation (coding, data entry, editing and correction)
2.5 Planned observation register
- 2.5.1 Overview
- 2.5.2 Object types, including derived object types
- 2.5.3 Object graph
- 2.5.4 Object/variable-matrixes, including derived variables
- 2.5.5 Data set descriptions
- 2.5.6 Derivation procedures (in complicated cases)

3 Completed data collection

3.1 Frame production
3.2 Sampling
3.3 Data collection
- 3.3.1 Communication with the data providers
- 3.3.2 Measurements, experiences of instruments
- 3.3.3 Interruptions/overcoverage, actions taken
- 3.3.4 Non-response, causes and actions taken
- 3.3.5 Editing and correction at data collection time

3.4 Data preparation (coding, data entry, *editing and correction*)
3.5 Production of final observation register
- 3.5.1 Treatment of interruption/overcoverage objects
- 3.5.2 Treatment of non-response objects
- 3.5.3 Treatment of partial non-response
- 3.5.4 Frequency counts of overcoverage, responses, non-responses etc
- 3.5.5 Completed derivations of derived objects and variables

4 Statistical processing and presentation

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

4.2 Population models
4.3 Computation formulae for estimations
- 4.3.1 Point estimations
- 4.3.2 Estimations of sampling errors (variance estimations)
- 4.3.3 Estimation/judgment of other quality characteristics

4.4 Analyses
4.5 Presentation and dissemination procedures

5 Data processing system

5.0 System overview
- 5.0.1 Verbal description
- 5.0.2 System flow

5.1* Subsystem description
- 5.1.1 Overview
  - 5.1.1.1 Verbal description
  - 5.1.1.2 System flow
- 5.1.2 Component descriptions
  - 5.1.2.1 Data sets
  - 5.1.2.2 Processes
  - 5.1.2.3 Other components

6 Log-book

Figure 6. *Documentation templet for a statistical survey and its production system.*

6 Confidentiality

Statistical data can only be made available to the users within the limitations of certain confidentiality restrictions. The most fundamental purpose of these restrictions is to preserve the data provider's confidence in the statistics producer's willingness and ability to ensure that data submitted to a statistics producer will be used for statistical purposes only. Among other things the statistics producer must be able to ensure that statistical outputs will not, thanks to the input submitted, directly or indirectly, enable a statistics user to associate sensitive information with the data provider or anyone whom the data provider would like to protect.

Statistical confidentiality can only be ensured by a combination of technical and legislative actions. Advanced statistical and mathematical methods alone will never be sufficient, however sophisticated they may be. This has been clearly demonstrated by massive research efforts during the last 25 years. Basically, statistical confidentiality is about confidence. A data provider, who does not trust a particular statistics producer, will not change his mind just because the statistics producer promises to apply a "perfectly safe" statistical method, if there were such a method (which there is not).

An adequate combination of technical and legislative rules for protecting the confidentiality of statistical data could be something along the following lines:

- It should be forbidden by law to use data submitted to a statistics producer for other than statistical purposes.

- Data submitted to a statistics producer for statistical purposes should be protected against sabotage, theft, and intrusion by physical and technical measures. Data that are associated with identified subjects (persons or organisations) must be handled only by authorised persons, "sworn in" by the statistical office.

- Statistical data must be anonymised (microdata) or aggregated (macrodata) before they can be distributed to users outside the statistical office. Anonymised microdata and aggregated macrodata must be checked by the statistics producer, so that they do not contain "obvious" disclosures of sensitive data for individual, easily identifiable subjects (persons, enterprises and other organisations). A disclosure is "obvious" if it does not require any conscious effort.

- It should be forbidden by law to make any conscious efforts to derive sensitive data about identified, individual subjects from statistical data.

- It should always be less attractive for a potential intruder, who considers all costs and benefits, to obtain information about identified subjects from protected statistical data than to obtain the same information from some other source.

- Statistical data that are not accompanied by adequate documentation (metadata) should be destroyed.

7 Experiences from Statistics Sweden

This paper has pointed to a number of problems and opportunities that need to be tackled by a statistics producer, who wants to make statistical data more available to a user, while satisfying restrictions given by scarce resources and the willingness of data providers to co-operate. The topics covered were:

- the "fuzzy" concepts of user-orientation and user-friendliness
- standard interfaces as instruments for simplicity and flexibility
- standard, "off-the-shelf" software components as instruments for speedy and inexpensive application development
- good quality metadata enabling the user to retrieve and process data independently of the producer
- technical and legislative measures for protecting the confidentiality of statistical data

Statistics Sweden is an example of a statistical agency, which has been working very actively in all these areas over the last three decades. In the late 1960's and early 1970's Statistics Sweden developed the TAB68 suite of high-level, non-procedural software products. These tools, which covered many important production steps, e.g. editing and tabulation, became extensively used at Statistics Sweden, first by non-programmers and then (after some initial hesitation) even by the programmers themselves. Many production systems are still heavily dependent on these software products.

After gaining important experiences from using the Canadian time series database system, CANSIM, Statistics Sweden developed its own AXIS system for making cross-sectional data as well as time series data available on-line to internal and external users. The system was put into regular operation in 1976, and it is still running successfully, although many users now demand data to be made available in many other ways than through relatively expensive and rigid main-frame communication. During the next few years the system will be phased out, and a new, client/server based system will be phased in. The new system is entirely PC based; it makes extensive use of standard interfaces, e.g. SQL and GESMES, as well as a wide range of "off-the-shelf" software products, favoured by internal and external users.

Figure 7 illustrates how the new statistical database system at Statistics Sweden is intended to co-operate with the survey-based production system within a client/server framework.

The new database system will make available a lot of aggregated macrodata (time series as well as cross-sectional), some anonymised microdata, and the metadata needed for efficient searching and responsible interpretation and analysis by external users. Microdata and macrodata will be stored in SQL databases. At a later stage object-oriented database management systems (OODBMS) and so-called on-line analytical processing (OLAP) products may be considered as alternatives or complements to SQL databases for certain types of usages.

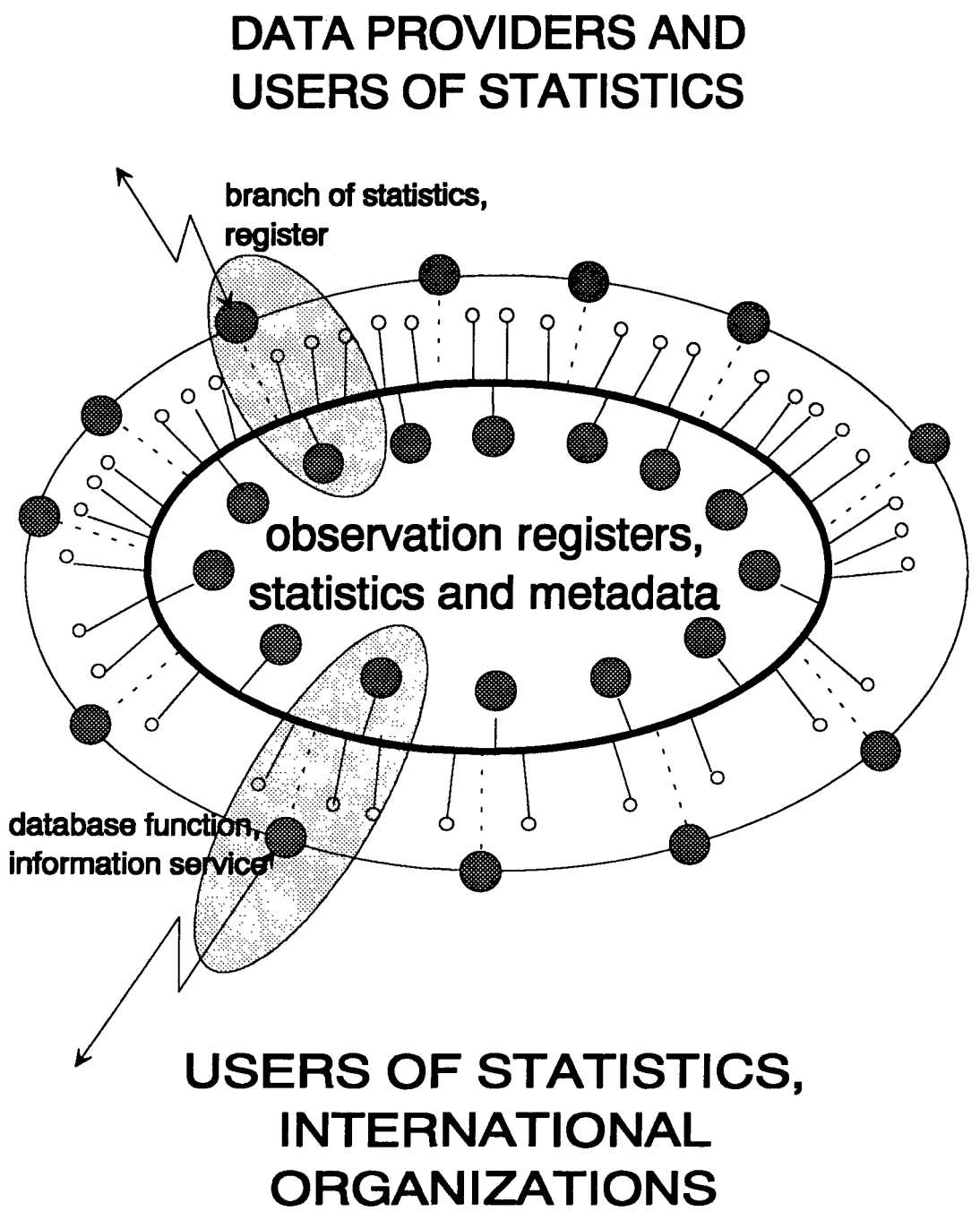

Figure 7. *Client-server architecture of a system of statistical information systems.*

The main sources of metadata will be survey documentations, following the SCBDOK documentation templet shown in figure 6 above, complemented by product overviews, quality declarations, and some other types of documentation, which are available for statistical products produced within the Swedish Statistical System. The bulk of metadata will be textual data with limited structuring. These data are most likely to be handled as a text database by free text searchers and document handling systems. A small but important part of the metadata are to be used for controlling the operation of various software products. These metadata need to be stored in an SQL database, so that they can be handled formally and automatically communicated and transformed between different software components inside and outside the database system.

The total size of the new statistical database, including metadata, macrodata, and anonymised microdata may turn out to be in the order of 100 GB.

Many different channels will be utilised for disseminating data from the new statistical database to the users, including self-service PCs in the premises of Statistics Sweden, available for external users, who want to down-load data and metadata from the statistical database to their own storage media, World Wide Web (WWW) databases, CD-ROM products, diskettes, etc.

As for confidentiality problems concerning statistical data (anonymised microdata and aggregated data with few contributors) the situation in Sweden has become dramatically improved for both users and producers as well as for data providers thanks to new legislation, which criminalises all attempts to derive identified data from statistical data. The particular paragraph about this in the Swedish Law on Official Statistics reads as follows:

"Official statistics must not be combined with other information for the purpose of finding out the identity of individual subjects."

In summary, on-going developments within the Swedish Statistical System provide good illustrations of the general principles that have been discussed in this paper. The practical results, which have been achieved so far, indicate that statistical offices will be able to meet the challenges from the users to make statistical data more available by means of modern technology, with due consideration to the interests of data providers and the public at large.

BIBLIOGRAPHY

Jackson, M. (1994), Problems, Methods, and Specialization, IEEE Software.

Johannesson, P. (1993), Schema Integration, Schema Translation, and Interoperability in Federated Information Systems, University of Stockholm.

Lebaube, P. (1991), EDI and Statistics - A Challenge for statisticians, Proc 48th Session of the International Statistical Institute, Cairo.

Malmborg, E. (1986), On the Semantics of Aggregated Data, Proc. Third Int. Workshop on Statistical and Scientific Database Management, Luxembourg.

Malmborg, E. (1992), Matrix-based Interchange of Aggregated Statistical Data, Proc. Sixth International Working Conference on Scientific and Statistical Database Management, Ascona, Switzerland.

Malmborg,E. and Lisagor, L. (1993), Implementing a Statistical Meta-Information System, In Eurostat Conference on Statistical Meta Information, Luxembourg, 2-4 Feb 93, also in Statistical Journal of the United Nations UN/ECE 2/1993.

Malmborg, E. and Sundgren, B. (1994), Integration of Statistical Information Systems - Theory and Practice, Proc. Seventh International Working Conference on Scientific and Statistical Database Management, Charlottesville, Virginia, USA.

Shoshani, A. (1982), Statistical Databases: Characteristics, Problems and some Solutions, Proc. 8th Int Conf on Very Large Data Bases.

Sundgren, B., (1973), An Infological Approach to Data Bases, Statistics Sweden, Urval Nr 7.

Sundgren, B. (1991a), Statistical Metainformation and Metainformation Systems, Statistics Sweden R&D Report 1991:11; also in Statistical Journal of the UN/ECE 2/1992.

Sundgren, B. (1991b), What metainformation should accompany statistical macrodata? Statistics Sweden R&D Report 1991:9.

Sundgren, B. (1992), Organizing the Metainformation Systems of a Statistical Office, Statistics Sweden R&D Report 1992:10; also in the documentation from the UN/ECE Work session on Statistical Metadata 1992 (METIS).

Sundgren, B. (1993a), Statistical Metainformation Systems - pragmatics, semantics, syntactics, In Eurostat Conference on Statistical Meta Information Systems, Luxembourg; also in Statistical Journal of the UN/ECE 2/1993.

Sundgren, B. (1993b), Guidelines on the Design and Implementation of Statistical Metainformation Systems, Statistics Sweden R&D Report 1993:4. ECE Work session on Statistical Metadata nov 1993, Revised versions 1994 and 1995.

UN/EDIFACT and Eurostat (1993), GESMES 93 Guidance to Users & Reference Guide (separate volumes), Eurostat, Luxembourg.

UN/EDIFACT (1994), Raw Data Reporting Message, Draft document.

Welke, R. J. (1994), The Shifting Software Development Paradigm, Proc. of the Baltic Workshop on National Infrastructure Databases, Vilnius, Lithuania.

R & D Reports är en för U/ADB och U/STM gemensam publikationsserie, som fr o m 1988-01-01 ersätter de tidigare "gula" och "gröna" serierna.

R & D Reports Statistics Sweden are published by the Department of Research & Development within Statistics Sweden. Reports dealing with statistical methods have green (grön) covers. Reports dealing with EDP methods have yellow (gul) covers.

Reports published 1995:

1995:1 (grön)	Asymptotic Theory for Order Sampling *(Bengt Rosén)*
1995:2 (gul)	PC-programvaror för estimation och statistisk analys *(Dan Hedlin)*
1995:3 (grön)	Vad användarna tycker om SCB - Rapport från två användarstudier med QSP *(Jan Eklöf, Mats Bergdahl)*
1995:4 (grön)	Bortfallsbarometern nr 10 *(Mats Bergdahl, Pär Brundell, Margareta Eriksson, Dan Hedlin, Monica Rennermalm, Per Sandgren)*
1995:5 (gul)	Statistical metadata - a tutorial *(Bo Sundgren)*

Tidigare utgivna R & D Reports kan beställas genom Ingvar Andersson, SCB, U/LEDN, 115 81 STOCKHOLM (telefon 08-783 41 47, telefax 08-783 45 99, E-mail ingvar.andersson@scb.se)

R & D Reports listed above as well as issues from 1988-94 can - in case they are still in stock - be ordered from Statistics Sweden, att. Ingvar Andersson U/LEDN, S-115 81 STOCKHOLM SWEDEN (telephone +46 8 783 41 47, telefax +46 8 783 45 99, E-mail ingvar.andersson@scb.se)