# The Impact of new Technology on Methodology and Organization of Statistical Data Processing

Bo Sundgren

SCB

R&D Report
Statistics Sweden
Research - Methods - Development
1988:15

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 1988:15. The impact of new technology on methodology and organization of statistical data processing / Bo Sundgren.
Digitaliserad av Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-1988-X101OP8815

# The Impact of new Technology on Methodology and Organization of Statistical Data Processing

Bo Sundgren

R&D Report
Statistics Sweden
Research - Methods - Development
1988:15

Published November, 1988
Printer STATISTICS SWEDEN, Dept of Research and Development, EDP Systems Unit
Publisher Staffan Wahlström
Questions answered by Bo Sundgren



ISSN 0283-8680

Printed in Sweden
Garnisonstryckeriet, Stockholm 1988

STATISTICS SWEDEN
Bo Sundgren

1988-08-08

# THE IMPACT OF NEW TECHNOLOGY ON METHODOLOGY AND ORGANIZATION OF STATISTICAL DATA PROCESSING

**Abstract**. During the last few decades we have witnessed a fantastic technological development, and an unbelievable reduction of the price/performance ratio for computers. Statistical offices have benefitted a lot from this development, by rationalizing their survey operations. We can now expect the new technology to be used more systematically in tasks requiring more "intelligence" than the basical computations and data handling operations in statistics production. Statistical design and EDP systems design are examples of such tasks. The new technology is also bringing about qualitative and structural changes, both inside statistical offices, and in the relations between statistical offices and its customers. Some effects, which are already becoming visible, are (i) standardization of technology, software, and methodology; (ii) decentralization of computers and computer-related resources; and (iii) integration of different types of tasks, skills, and competences. In-house software development is being critically examined by several statistical offices, and it is an open question whether we shall see new generations of generalized software products specially destined for statistics production, and what they will look like. International cooperation in software development between statistical offices could be an attractive alternative. Outside statistical offices the new technology will offer new possibilities to the users of statistics, and the users will demand better service from statistics producers. This may call for some rethinking and reorganization within statistical offices.

THE IMPACT OF NEW TECHNOLOGY ON METHODOLOGY AND ORGANIZATION OF STATISTICAL DATA PROCESSING

Table of contents


1 The historical development and today's situation . . . . . 3

2 Different types and levels of computer support 4
2.1 Routine applications vs "intelligence" . 4
2.2 Survey operation vs survey planning, administration and evaluation vs strategical planning of the tasks and organization of a statistical agency . . . . . 5

3 Continued rationalization of survey operations through standardization, decentralization, and integration . . . . . 6
3.1 Standardization of technology, software, and methodology . . . . . 6
3.2 Decentralization of computers and computer-related resources . . . . . 9
3.3 Integration of different types of tasks, skills, and competences . . . . . 10

4 Computer-assisted design of statistical surveys and statistical information systems . . . . . 12
4.1 Statistical design . . . . . 12
4.2 Systems design . . . . . 13
4.3 Knowledge-based methods . . . . . 15

5 The next generation of statistical software . 17
5.1 International cooperation: The UN/ECE Statistical Computing Project (SCP) . . 17
5.2 Some desirable properties of the software 18
5.3 A proposed architecture for the software 19

6 The future architecture of statistical information systems . . . . . 23
6.1 Is there a need for a statistical office any longer? . . . . . 23
6.2 User needs . . . . . 23
6.3 Needs for rethinking? . . . . . 24

## 1 The historical development and today's situation

The impressiveness of the technological development and the capacity and inexpensiveness of today's computers is well-known and need not be repeated here. It may be enough to point to the fact that each one of the personal computers standing today on the desks of individual staff members of a statistical office has approximately the same capacity as the whole mainframe of the statistical office 20 years ago, but at a fraction of its price. I think that it is also fair to say that most of these personal computers are not just standing there on the desks as a kind of status symbol, but they are being used very productively in the work of the statistical office.

It may be interesting to contrast this actual situation with some forecasts about the needs for computers that were made a few decades ago, when this technology emerged and became practically available. At that time some of the leading computer experts in Sweden seriously believed that one computer would be more than enough for all the needs of our country for the forseeable future. Similar judgments were made in other countries. This type of prognoses is all the more remarkable, since if we look at the functionality (rather than capacity and price) of computers, the development has not been that dramatical. On a low, technical level, computers do essentially the same things now as then, only so enormously much more efficiently. Thus in principle the computer experts of 30 - 40 years ago should have been able to prognosticate a little more accurately the potential of computer technology. But they did not. The mistake they made was that they considered only one narrow category of applications, mathematical computations, and even for that type of application they did not have the imagination to foresee the explosion of needs that would appear, once the technology was available on a large scale and at a minimal price.

One thing I want to say with this is that from the very start of computer history, we seem to have been lagging behind in our ability to fully appreciate the application potentials of computer technology, and to actively plan for the most constructive usage of these potentials. This seems to be true also for statistical offices. We are eager to acquire the most recent computer technology, but, in my opinion, we are far too often spending too much of our resources just to move the same old applications between different generations of technologies, rather than actively developing new applications, new methodologies, and new ways of performing the overall tasks of a statistical office, based on a little more imaginative, long-term, strategical judgments about the future availability of computer technology. We also

seem to neglect the drastical changes in cost relationships that are taking place all the time between hardware, software, and personnel resources, and to fail to ask ourselves more explicitly now and then whether not quite new mixtures of these production factors would be more optimal. For example, at Statistics Sweden, until recently, the acquirement of a personal computer had to be formally approved by the Director General of our agency, whereas the employment of a secretary, an investment that is roughly 1000 times bigger than the purchase of a PC, and with a 10 times longer "write-off period", could be decided on a lower managerial level.

## 2 Different types and levels of computer support

In a statistical agency there is a wide spectrum of possibilities to use computers. Some of these possibilities have already been exploited to a great extent, whereas others are at best in the prospecting stage. In order to discuss the potential of modern technology in statistical work, we need a basic structuring of that work. I have chosen to use two alternative structures. One is a classification of statistical tasks into those which are of a more or less routine character, and those which seem to require more "intelligence" of one sort or another. According to the other classification we make a distinction between the statistical operations as such, on the one hand, and the control of statistical operations on the other, where "control" includes planning, administration, and evaluation. Among the control tasks we may again distinguish between those which aim at individual statistical surveys, and those which have a whole statistical system as their object, for example, the whole statistical system of a country.

### 2.1 Routine applications vs "intelligence"

The vast majority of computer applications in a statistical office today are of a rather routine nature. Data are entered, edited, sorted, counted, and presented in a fairly straight-forward way. The computations are not always very complicated, but the volumes of data are sometimes quite large. The computer is little more than a pedantic, incredibly efficient book-keeper, who makes no errors. Nevertheless, this has turned out to be good enough to save large amounts of money for statistical agencies.

However, the challenge that is now facing us is whether we can start using computers in a more "intelligent" way. So far we have been very successful in multiplying the human being's capability to move and sort data, and to count them, and to eliminate the human tendency to

commit errors in those operations. But can we also use the computer as an amplifier of the human intellect in statistical work? Without exaggerating the possibilities of disciplines with fancy names like "artificial intelligence" and "expert systems", I think that there are many good opportunities of using **knowledge-based methods** in statistics production. We shall return to this issue in chapter 4.

## 2.2 Survey operation vs survey planning, administration and evaluation vs strategical planning of the tasks and organization of a statistical agency

As was indicated above, we may look at the work of a statistical agency on three different levels. On the first, basic level, we have the actual statistical operations, making up the operational parts of a statistical survey. We all recognize the traditional, serial flow of tasks that have to be performed, when we conduct a survey: data collection, coding, editing, data transformation, aggregation, tabulation, graphical presentation, analysis, publication, distribution. Still existing, unused potentials for development of the computer-support in this area will be discussed in chapter 3.

On the second level, we **control** the different steps in the survey, and the survey as a whole, by means of design and planning, administration, and evaluation. The statistical design includes the establishment of a frame and a sampling strategy, if any, and the EDP design includes systems analysis, data modelling, and programming. The potentials for improved computer-assistance in these tasks will be treated in chapter 4.

Finally, on the third level, we look upon a statistical system as a whole as the object of control. The statistical system under consideration may be the statistical information system of a country, or a part of such a system that is managed by a particular statistical agency. Even though such a system will be based upon a number of statistical surveys, it will also contain other components, and it is true for statistical systems as for other systems that the whole system should be something more than just the sum of its parts. In chapter 6 we shall discuss in what ways a statistical information system can be something more than the sum of the surveys that it contains, and how a systems approach based on modern technology can help the supersystem to fullfil its purposes as effectively as possible.

## 3 Continued rationalization of survey operations through standardization, decentralization, and integration

One of the principal messages of this paper is that we should actively look for possibilities to use modern technology in new areas and new aspects of statistics production, rather than just being busy moving "the same old applications" from one computer generation to the other. On the other hand, we must not neglect possibilities to do "the same old things" in a much better way by applying new technology and new methodology to the traditional tasks of the operation of a statistical survey. I will use three slogans to describe what needs to be done: standardization, decentralization, and integration.

### 3.1 Standardization of technology, software, and methodology

Computers are now so cheap, and people so expensive, that it is very rare that it is really worthwhile to aim at maximum technical efficiency in the design of a computerized information system. Naturally, in a big system, with large volumes of data to handle, and with a heavy traffic of man-machine interaction to take care of, it may sometimes be necessary, or at least clearly cost-efficient, to mobilize technical inginuity in order to eliminate potential bottle-necks and speed up response times, or save storage and processing capacity. But there are two points to be made here. One is that there are not so many systems of this nature, not even in a statistical office with its large data bases. Most statistical surveys are small or modest in size, they are processed rather infrequently, and response time requirements are often quite moderate in comparison with those of many commercial on-line systems of administrative character.

The other point I wanted to make in this context is that even in those few cases, where technical optimization is really optimal, from an executive point of view, it is usually not executively optimal to technically optimize the system as a whole, but only some limited part or aspect of it.

As a consequence of these observations, when it comes to the rationality of technical optimization, the burden of proof should always lie with those who claim that it is necessary. The default solution should always be a simple, straightforward, standardized solution.

What I have just said may seem to be so obvious that it need not be said. Unfortunately, this "obvious truth" does not always seem to be well understood or widely

accepted in statistical offices. Even if we theoretically accept its validy, we seem to neglect its consequences when it comes to practice, not least when it comes to managerial practice. This seems to be a case where some active **delearning** of an old, puritan habit (not to waste any type of material resources under any circumstances) should be exercised among all those involved: programmers, systems analysts, and managers, to the benefit of the overall economy and usefulness of the systems designed.

In other words, a systems analyst or a programmer should never be allowed to deviate from the standards, set by the top management of a statistical office, without the hearing of a responsible manager, and a responsible manager should never accept a deviation from the standards without lengthy arguments, based on good documentation, the contents of which he or she fully appreciates. Violation of these rules should be regarded as a serious fault by internal and external auditors.

The proposed managerial rule assumes that there is a well worked-out standard, or policy, controlling important and tangible aspects of information systems design in the statistical agency. Ideally this policy should be well integrated with a theoretically sound methodology for systems design and implementation, so that the rules of the policy will be more or less automatically followed by anyone who uses that methodology, and its accompanying working tools (cf section 4.2). Objects for standardization are hardware, software, interfaces, and the methodology itself, including documentation rules.

**Standardization of hardware and operating systems**

For the time being the standardization of hardware and operating systems is not really a major problem, since the industry has (just by chance?) solved it for the users by establishing very strong **de facto** standards: IBM-compatible mainframes, IBM-compatible microcomputers, possibly UNIX. However, it should be noted and kept in mind that this good situation is not at all the result of some explicit action on the part of the users. Thus the scene may change rather quickly back to the more traditional lack of standards.

**Software standardization**

The most important phenomenon in the software area is that it seems to have become finally accepted that users should not develop their own application specific software. Instead they should use generalized software and, if necessary, customize these products for their specific applications. As a result, the number of application

programmers should decrease rather drastically in statistical offices, if it has not already started to do so. Some statistical offices have even started to question their own development of generalized software. Thus it seems that we are not far from questioning the whole professional category of programmers as an identifiable group of specialists within statistical offices. We shall return to this controversial issue.

It should be noted that when we talk about standardized software, or generalized software, in a statistical office, there are two important subcategories. One type of generalized software is the product which has been developed for particular statistical tasks, or functions that are typical for statistical offices. Such **generalized statistical software**, or **specialpurpose generalized software** has often been developed by the statistical offices themselves, or by institutes doing statistical research. There is another type of generalized software that we may call **multipurpose generalized software**. Such a product has typically been developed for general, commercial purposes, and not in particular for statistical applications, although they may also be useful in statistical environments. Data base management systems are a good example.

**Standard methodology**

Today most major computer-using companies and organizations in Sweden have adopted some standard methodology for the development and maintenance of information systems. The methodologies are called **systems development models**, and they say something about what tasks should be performed during systems development, in which order they should be performed, what concepts to use during analysis, how to visualize the concepts and the results of the analysis, which rules and standards are to be followed, which documents to prepare, etc. Typically there is one standard, or at least one variation of a standard per company. However, some kind of **de facto** standard seems to be emerging, even between the companies, including such features as separation between infological (contents-oriented) and datalogical (technique-oriented) phases, emphasis on conceptual modelling according to some "three-concepts-methodology" (objects, properties, relations [OPR], or entities, attributes, relations [EAR]) during the infological phase, and emphasis on the relational data model during the datalogical phase. As a complement to the state-oriented modelling of concepts and data, the systems development methodologies often contain flow-oriented modelling techniques and other methods for clarifying dynamical aspects of the system.

One important aspect of systems development methodologies is that they may prescribe **standardized interfaces**, for example standard formats for the storage and communication of data and metadata between different parts of a system, and standardized user interfaces in general, or standard syntaxes for user languages in particular.

The systems development methodology of Statistics Sweden is called the **SCB model** and has been in use for over a decade. It has also been the source of inspiration for the systems development models of several major (non-statistical) companies and organizations outside Statistics Sweden.

## 3.2 Decentralization of computers and computer-related resources

An obvious organizational consequence of the changes in cost relationships should be that there is no longer the same need for centralized control of computer-resources as there used to be. We are now able to buy such resources in small pieces and at a very low cost per piece. Thus there is no need for centralization for the reason of sharing expensive, indivisible, and scarce resources. Instead we can integrate the decisions concerning computer resources with other important decisions in the statistical office and try to develop the same type of "balanced decentralization" of decision-making as in other areas, letting those responsible for a statistical survey take as full responsibility as possible for all types of resources needed for the design and operation of the survey, not treating computer resources in any special way.

In Sweden we have taken some important steps in this direction. The Government has initiated a process where those agencies which have enough competence and experience in EDP are allowed to take most computer-related decisions without having to ask any other agency or the minstry of finance. A condition for this freedom is, of course, that the agency is able to handle all decisions within its given budget.

Within Statistics Sweden we are also trying to treat EDP-related decisions, not separately, but integrated with other decisions. We have established an EDP policy, which will of course be updated from time to time, and within this policy, each department is authorized to take its own decisions, as long as they are within the budget of the department. Thus, as far as possible, all types of costs - for hardware, software, and personnel, for mainframe-related resources and for micros - are measured in "the same kind of money".

In the area of application systems design and programming we started the decentralization process already five years ago. The responsibility and the personnel resources for these activities were then removed from the central Systems Department to the different subject matter divisions. The remaining parts of the Systems Department were merged with some other development functions into a newly formed division for research and development and retained the responsibility for such functions as development and maintenance of generalized software, research and development in the area of statistical data processing, and EDP training.

Within each subject matter division the decentralization process has continued more or less rapidly. In some cases there are still some relatively big pools of systems analysts and programmmers, and in some other cases the decentralization process has continued down to the level of individual subject matter units and surveys.

One good effect of this decentralization is that the manager of a particular statistical survey has now much more complete overview, knowledge, and control of his/her product and all types of resources that are needed, assuming of course that the manager has the capacity and willingness to make use of these opportunities. On the other hand there is naturally a risk that the statistical office as a whole will fall apart into a large number of small, uncoordinated survey based organizations. In order to prevent this, a number of specialized "councils" (among others one for EDP) have been created for giving specialized advice in policy matters etc to the top management and the Director General of the office.

In my opinion this decentralization process has by and large been successful and necessary and will continue in the future. Hopefully this will among other things lead to a better integration of different methodological aspects of statistics production, including the integration of statistical methodology and EDP. We shall return to this issue in the next section.

### 3.3 Integration of different types of tasks, skills, and competences

Integration of different tasks, skills, and competences is the other side of the "decentralization coin". The effects of this integration are becoming visible throughout the organization. Managers are losing their personal secretaries and administrative assistants, who find themselves replaced by personal computers, networks,

and office information systems. Subject matter statisticians are taking care of application development and maintenance without the assistance of systems analysts and programmers, and they do some of their own publishing without having to rely on typographers. Interviewers take over data entry and data editing tasks.

Integration of these and similar types have been enabled and facilitated by the decentralization of computers and computer-related resources. The integration has a number of good effects, including job enlargement, shorter communication and decision paths, less administrative overhead, and a more clear division of responsibilities. On the negative side there is a risk of "happy amateurism" replacing competent and efficient professionalism, and of isolationism and self-conceit in the relatively small and independent organizational units. However, on the whole the positive effects seem to outweigh the negative ones, and there seem to be more staff members who feel they have gained from the development than who feel they have lost. Even some of those who have lost responsibilities and empires welcome or accept the development as being basically sound and find new roles in the organization relatively quickly.

Naturally, there will always be a need for good specialists in several fields of competence in a statistical office. The on-going technological development only eliminates a need for centralization and functionalization that was based on the indivisibility of large, expensive computers, and on the relative scarcity of systems analysts and programmers. However, the development has also raised the question, whether it would not be rational to aim at a higher degree of integration between the disciplines that are of relevance for statistics production. The division into specialties that exist at present in statistical offices by and large reflect the university organization. For a statistical office it would clearly be fruitful with a closer cooperation between statistical methodologists and computer specialists. Such an integration would be even more important, if it could also induce a change at the universities, making statistics production **per se** an established and respected area of academical research and qualification.

## 4 Computer-assisted design of statistical surveys and statistical information systems

No doubt, the next wave of computerization in statistical agencies will concern the environment of the statistical operations, rather than the statistical operations themselves. By and large we have already computerized the basical data handling and computation that is going on in the processing of a statistical survey. Now it is time to seriously consider a massive computerization of the tasks involved in the control of a statistical survey, tasks like planning and design of the survey from a statistical point of view, design and construction of the production system (which is in itself for the most part a computerized system), administration of the production activities, and evaluation of the performance of the statistical survey. Some of these tasks, and their computerization potentials, will be examined in this chapter.

### 4.1 Statistical design

In a statistical environment it is, of course, a well recognized fact that the statistical design is a most important activity in the planning and execution of a statistical survey. I am also sure that most statistical methodologists have already accepted the computer as an efficient calculator and tool in their design work. An interesting question is whether the computerization of the statistical design can take place in a more systematical and goal-driven way. The answer to this question can only be given by the statisticians themselves, in cooperation with computer specialists.

My personal belief is that statistical methodology, supported by an intelligent use of computers, could produce another round of rationalization of statistics production of maybe the same order of magnitude as the computers alone have already accomplished. It seems that non-statisticians sometimes think of statistical methodology as a "necessary evil", which has to be there in order to ensure reasonable quality of the results of a statistical survey, and to protect against conscious and unconscious misuse of statistics. Naturally this is an important function of statistical methodology, but statistical methodology in combination with modern technology could also be a very powerful, active force in the rationalization and cost-saving in statistics production.

Let me mention a couple of examples. In a major Swedish household survey, a team of competent statistical methodologists could, with the help of computerized analyses and simulations, propose a more efficient sampling and

stratification strategy. The results from this exercise could be used for an improvement of the quality of the estimates produced in the survey. Alternatively the sample size could be reduced by some 50%, thus halving the data collection costs, a major budget item (about 2 MSEK) in this survey.

Another type of example is the significant gains that can often be made by a well designed combined usage of sample surveys and administrative registers. In Scandinavian countries this design could eliminate the need to take censuses, thus saving vast sums of money. An interesting characteristic of the approach is that it seems to be able to improve the quality of both the register and the survey, at the same time as it reduces costs.

Naturally a statistical methodologist could make this list of examples much longer and more precise. However, statistical methodologists do not seem to advertise and market the rationalization power of their methodological tools in this way very often. As a result, it seems to me that too many subject matter statisticians too often turn their interest to the latest novelties in computer technology, rather than seriously investigating some of the rationalization potentials offered by computer-supported statistical methodology.

### 4.2 Systems design

The growing usage of generalized software instead of tailor-made application programs has greatly improved the efficiency in systems construction and maintenace. However, the use of generalized software does not decrease the need for a good systems design. Nor does it in any significant way reduce the amount of work that has to be done by competent specialists during earlier design phases. It may seem surprising that exactly those specialists, who are themselves responsible for the design of so many computerized systems, do not to a greater extent use computerized systems to support their own efforts. After all we have since long got used to concepts like Computer Assisted Education (CAE) and Computer Aided Design and Manufacturing (CAD/CAM).

However, right now something is happening in this area. A new acronym has been coined, CASE, standing for Computer Assisted Software Engineering, or **Computer Assisted Systems Engineering**. The CASE tool-boxes contain software instruments supporting the different working steps in systems development models. One problem is that each organization has its own systems development model, and there may not be any particular CASE tool-box on the market, which perfectly matches the needs

implied by that model. The organization is then left with the alternatives of either having to change its systems development methodology, or to develop its own CASE tool-box.

As I mentioned in section 3.1, there is a process of **de facto** standardization going on in the area of systems development models. This in combination with the possibilities for an organization to acquire and customize a **CASE shell** (cf expert system shells), rather than having to accept all the details of a completely ready-made CASE tool, should help to solve the problem.

For a statistical office this development should be of great interest. While waiting for an adequate CASE shell to appear on the market, the statistical agency could itself undertake a number of relatively simple steps to improve the computer support in systems design.

One obvious, but important step is to develop an interactive tool for the creation and maintenance of systems documentation.

Another, related development step is to have a tool that automatically transforms and communicates metadata between different software products.

Furthermore, it could be questioned whether the designer/user should at all have to bother about more or less unimportant technical differences between different software packages, or even with the selection of a particular software product (rather than another one) in the first place. Ideally the designer/user should only have to specify the function (for example tabulation) that he or she wants to be performed, and then the systems development tool should automatically select (or propose) a software product and generate a complete application on the basis of metadata from the documentation system and some input from the designer/user, expressing his/her preferences on certain matters.

A documentation and systems development tool approaching the above-mentioned ideals has been developed at Statistics Sweden. It is called the **CONDUCTOR** and is running on the mainframe at present. It speeds up the work even of experienced systems analysts, and it makes it possble for people who are not computer professionals to get their own applications "in the air", provided that they have a relatively simple problem and/or an adequate understanding of the early, contents-oriented phases of the systems development model.

## 4.3 Knowledge-based methods

The term "knowledge-based system" is often used today as a more humble way of saying "expert system", which is in turn intended to be more down-to-earth than "artificial intelligence". Anyone who is not familiar with this jargon might rightfully question why we should suddenly need to start talking about using "knowledge-based systems" and "knowledge-based methods" in statistics production. (Have we not always used methods based on knowledge? What other methods could there be?)

It is true, hopefully, that statistics production has always been based on knowledge, but typically this knowledge was not stored outside human brains, and if it was, it was usually stored on paper, separately from the computerized files, containing the data that were processed in accordance with the knowledge. And finally, if the knowledge was to some extent represented in a computer, it was usually stored implicitly in the programs.

In contrast, knowledge-based computer methodology assumes that

* the knowledge used in different parts of statistics production is (at least partially) computerized;

* the knowledge is organized as **facts** and **rules** in a so-called **knowledge-base**, which is handled in accordance with data base principles;

* the exploitation of the knowledge is actively computer-supported.

Thus, even though it is controversial, it must be admitted that one goal of applying knowledge-based methods to statistics production is to capture at least some small parts of the knowledge, which has up to now been regarded as inseparable from the statisticians who are in possession of the knowledge, and make it available to computers and to users of computers.

We should rightfully question to what extent this goal is a realistic one, but I think that we are not in a position to reject these ideas and proposals categorically. We must realize that we have only seen the beginning of a data explosion in society. Technically, anyone will soon be able to produce "statistics" from these data. But how can we prevent misinterpretation and misuse of this statistics production? The best thing would be, of course, if every amateur statistics producer would

seek the advice of a competent and experienced statistician. But even if there were statisticians in such abundance that this would be a realistical possibility, I am not sure that most users would follow this path, and I am not sure that the competent and experienced statisticians would appreciate to spend 99% of their time giving routine advice on routine statistical problems.

I think that if we reason along these lines we can rather easily agree on a justified, desirable and realistical role for knowledge-based computer methodology in statistics production. In this perspective the knowledge-based systems is a natural step in the development, following the metadata systems and interactive user interfaces (cf the above-mentioned CONDUCTOR system) that we have already put into productive use, and which are appreciated by most of us.

Moreover I am rather convinced that the efforts to develop knowledge-based computerized systems for statistics production will generate some very good side-effects, even if the more ambitious goals should not be reached so easily. I think that we all agree that a statistical agency has its most important asset in their staff members, and in the competence and knowledge they possess. We have a big problem to maintain this knowledge capital when specialists retire, or when budgets are cut. Systematical documentation of the knowledge (called **knowledge acquisition** in the jargon of expert systems) could alleviate these problems and provide excellent instruments for in-house training.

## 5 The next generation of statistical software

When starting a discussion about the next generation of statistical software, the first relevant question is, whether there will be a next generation of statistical software, at least if we are thinking about software products developed by statistical offices themselves. Some statistical offices have already started to question the need for programmers and in-house software development.

I am sure that I am not the only one who would hate to see a statistical office without some competence in advanced software development. A disengagement in this area is probably an irreversible process, and it will have negative side-effects. For example, it may turn out to be difficult to evaluate, select, install, and adjust commercial software to the specific needs of a statistical office, if the office does not have a critical mass of competence in software development.

On the other hand it must be admitted that it will be increasingly difficult for statistical offices to justify glamourous software development projects of the costly type that we used to launch in the past, and which we sometimes (but not always) managed to complete and implement successfully.

There are some actions that can be taken in order to come to grips with this difficult situation. One is to establish a policy and basic architecture for software development, ensuring important features like cohabitation possibilities between commercial packages and in-house developed software components, modularity and incrementality in software development, and portability between different types of computers.

Another possible action, which should be combined with the first one, is to rely more on international cooperation between statistical agencies. I will use some experiences from the UN/ECE Statistical Computing Project (SCP) as a basis for the discussion of these matters.

### 5.1 International cooperation: The UN/ECE Statistical Computing Project (SCP)

SCP is an acronym that symbolizes a cooperation effort in the area of statistical computing between the countries of the UN Economic Commission for Europe (ECE), including the European countries, Canada, and the United States of America. SCP has been going on in various forms and with various themes of cooperation throughout the 1980s. It started as a project (SCP-1), supported to some extent by the UNDP, then it became a programme

(SCP) under the Conference of European Statisticians, and since about a year and a half it is once again a project (SCP-2), supported by the UNDP. Whereas SCP-1 was basically mainframe-oriented, SCP-2 should pay special attention to the growing use of micro-computers in statistical offices.

The substantial work in SCP-2 has been organized into six Joint Groups. They are:

* the Joint Group on Software Evaluation
  (lead country: Hungary)

* the Joint Group on Communication
  (lead country: France)

* the Joint Group on Implementation Strategy
  (lead country: Poland)

* the Joint Group on Statistical DataBase Management
  (lead country: Sweden)

* the Joint Group on Data Editing
  (lead country: Yugoslavia)

* the Joint Group on Table Generation
  (lead country: German Democratic Republic)

The work of the Joint Groups is monitored by a Task Force, consisting of the lead countries of the Joint Groups, and the Task Force reports to a Steering Committee consisting of all countries participating in SCP-2. The three last-mentioned Joint Groups in the list above are actively engaged in software development in their respective fields of interest.

## 5.2 Some desirable properties of the software

In an attempt to amplify the total effect of the software development going on in the different Joint Groups of the Statistical Computing Project, the author of this paper was asked by the Task Force of SCP to come up with some concrete coordination proposals. Ideally this effort should result in

- a description of a unified design approach and software architecture to be shared by all the Joint Groups in the continued software development

- a proposal for a unified way of handling metadata in SCP software products

- a proposal for ensuring easy import/export

of data between different software products (SCP software, commercial packages, homemade products etc) and between different computers

- a proposal for ensuring portability for SCP software between different types of computers (micros/minis/mainframes)

- a suggestion of steps to be taken to facilitate the "marketing" of SCP products as members of one and the same software family

- a tentative, synchronized plan of activities to be carried out by the individual Joint Groups in order to fullfil the common goals of the SCP software development

Some of the proposals from this mission will be presented in the next section.

### 5.3 A proposed architecture for the software

The report put forward as a result of the coordinative effort mentioned above includes the following proposals:

[1] All SCP software development should be based on a data base oriented model of statistics production. The elementary operations of this model should be carefully defined, and the definitions should be based on a logically stringent, functional analysis of the typical major functions in statistics production. All intermediary operations in a production chain should use and produce data base objects of one and the same type: flat files, or relations in the sense of the relational data model. Thus in a mathematical sense, the operations would constitute an algebra over this type of data structures. This conceptual basis for the SCP software development will ensure modularity, simplicity, combinability, and incremental developability.

[2] In addition to portable and well integrated basic software components for editing, data base management, and tabulation, the SCP software package should contain a user-friendly Systems Development Environment (SDE), consisting of an Interactive User Interface (IUI) and, if possible, some tools based on the principles of Computer-Aided Systems Engineering (CASE) and knowledge-based methodology (expert systems). Among other things, the IUI should help the user to overcome any differen-

ces in the user languages that may exist between different SCP software components, and that will certainly exist between the SCP software package and other software products that the user may want to combine with the SCP software. Thus the IUI should help to standardize the user interface despite inevitable differences between software products. On the other hand, the IUI could also be used to individualize (customize) the user interface to fit the particular needs of a particular user group, or a particular statistical office.

[3] The uniformity of the SCP software products on the conceptual level should have a natural counterpart on the technical level. An algebra of operators working on standardized data structures (flat files and relations) has already been successfully implemented, and these principles should be generalized and applied to the other SCP software projects as well. Here a processor concept is proposed as the software technical counterpart to the algebra operators. Each SCP software component should be designed in terms of processors, and processors with identical or similar tasks should be standardized, and implemented only once. Thus identically the same processor could be used in several parts of the same software component, and in several software components.

[4] The processor language, controlling the operations of the processors, could sometimes also be the user language. For example, this is the situation in the case of the Base Operator System. For more complex functions like editing and tabulation, there is good reason to have a special user language that is mapped (translated) into the processor language.

[5] In order to standardize and facilitate the accessing and communication of data within processors, between processors, between SCP software products, and between SCP software products and the outside world, a common Relational Access Method (RAM) is proposed to be implemented. This access method should be used by all processors in all SCP software products for the reading and writing of data. RAM should consist of a set of independent macros, which can be included in the processor modules. In order to connect an external software package to RAM, one will have to

develop the appropriate read/write macros etc for the particular package, but this will be a relatively minor task.

[6] Similarly, in order to standardize and facilitate the handling of metadata, including - as far as possible - automatical transfer and transformation of metadata between processors, between SCP software products, and between SCP software and external packages, a common Metadata Management System (MMS) is proposed to be implemented. MMS should be used by all processors in all SCP software, and it should consist of a set of macros for the reading, writing, updating, deleting etc of metadata. Using such a set of macros, it would be quite easy to support different types and formats of metadata without any changes in the processors. However, it should be noted that this is a proposal for a standardization of the handling of metadata, not for a standardization of all kinds of metadata, which does not seem to be a realistical objective at present.

[7] In order to ensure maximum portability of SCP software between different categories of computers and operating systems, the programming language C is proposed to be used in all software development. The portability should (with priority 1) be certified for IBM PC compatible micros under PC/DOS and MS/DOS, for IBM 370 compatible mainframes under OS/MVS and VM/CMS, and for the operating system UNIX.

[8] The design, implementation, and documentation of all SCP software products should cover the following items, in the following order:

* systems analysis and formal description of the particular statistics production funtion under consideration

* development of a reference manual for the end-user language

* module specification in pseudo-code

* development of a systems manual for the software system under consideration, containing a description of

  - the logical program structure

- the processor language statements that are used and produced by the software

- the possibilities to write tailor-made exits

- other features for an advanced use of the software

* coding in C

* preparation of

  - a user's guide, based on pedagogical examples

  - an installation guide

  - an installation tape and installation diskettes, containing the software and its documentation

  - additional documentation if necessary

[9] In applicable parts, the design, implementation, and documentation of common SCP software components, like the processor language, the Relational Access Method (RAM) and the Metadata Management System (MMS), should cover the same items as listed above.

**Some comments to the proposals**

The proposals presented above are on a relatively high level of ambition, and of course it remains to be seen to which extent they will be accepted and realized by the SCP Joint Groups. In particular it may be noted that the proposals assume that the software development will be carried out in the programming language C. An alternative could be to base most of the software development on program generation techniques and/or some commercially available, portable, general-purpose software product like the database management system ORACLE with the standardized interface SQL. Most of the other proposals in the list above would be relevant anyhow, but naturally they have to be reinterpreted to some extent.

# 6 The future architecture of statistical information systems

## 6.1 Is there a need for a statistical office any longer?

It was noted in chapter 3 of this paper that the technological development has alleviated one restriction on statistics production that used to exist: the necessity to share scarce and expensive computers and computer-related resources. This has started a decentralization process. Are there any natural limits to this development, or will the decentralization stop only when the statistical office has been dissolved into a number of separate statistical surveys? In other words: will there be a role to play for a statistical office as a strong, independent organizational entity in the future, or will the statistical surveys be taken care of by other governmental agencies?

Personally I do believe that statistical offices have an important role to play in society, quite regardless of the decentralization possibilities that the technological development is now offering, but I think that we need to ask those critical questions indicated above. Others will do it.

Statistics production in Sweden was centralized into its present form in the early 1960s. The need to rationalize efficiently by means of centralized computer technology was then a major reason for centralization. But there were also others with at least the same dignity. One was the belief that only a strong, central statistical office could afford to maintain a powerful methodological development of high quality and enough quantity to form a "critical mass". Another reason for centralization was the needs for coordination and integration of individual surveys into statistical information systems, based on unified conceptual models like the system of national accounts and socio-demographical and socio-economical models.

Until recently the technological arguments for a centralized statistics production have been so widely accepted that we have not had to use the other, more sophisticated arguments. Maybe, as a consequence, we have not been so active in the areas of methodology and integration as we should have been.

## 6.2 User needs

The needs for coordination and integration are deeply and directly founded in some strongly felt user needs. For example, the statistics customers of Statistics Sweden are rightfully irritated when they have difficul-

ties to locate and interpret the statistical data that they are looking for, and they are equally rightfully irritated when they have to go to several places in the organization in order to get all the data they need, instead of getting everything in one place, including some advice about how to combine data from different sources.

Furthermore there is a growing number of rather advanced users of statistics, with more or less sophisticated models and hypotheses that they would like to try on official statistical data, and sometimes combine with their own data. Due to the technological development these users will always have access to powerful computer equipment of their own, and they have a good understanding of the possibilities offered by modern technology. If these users are not well served by the statistical office, they will exercise all the rights that they may have to obtain statistical data in rather "raw" form from the statistical office, and use them together with their own data, software, and models in **data laboratories** that they build and run independently of the statistical office. If a statistical office wants to be successful in this competition it must be active, imaginative, and flexible, and it must use its relative advantages in methodological competence, and coordination and integration possibilities.

### 6.3 Needs for rethinking?

The statistical survey is the basic building block in many statistical organizations. I have pointed out that the on-going decentralization will further strengthen the power and control in the hands of individual statistical surveys. From a managerial point of view, this development is good in many respects. It clarifies responsibilities within the organization, and the person who is in charge of a statistical survey will not have so many others to blame, if something goes wrong.

On the other hand there are those user needs discussed in the previous section, which call for other organizational solutions. In order to make it easy to locate and interpret data, all statistical data of any importance must be well documented, and they must be documented in the same way from survey to survey. Thus all statistical data must be accompanied by appropriate **metadata**. The metadata must be computerized, and like the data themselves they must be organized in accordance with **uniform database principles**. Many statistical offices have since long been trying to implement these ideas in different ways, but the results are not always encouraging. In today's competitive situation it will not be enough for some surveys and some departments to

be engaged in this work. Instead a systematical implementation throughout the organization may be a question of vital importance for the statistical office as an independent entity.

Databases and metadata will not be enough to serve the needs of the users. With all data and metadata easily available, they are certain to **combine data from different sources**, that is, data emanating from different surveys and other sources, like administrative registers. And they will make such combinations whether we approve of it or not. Traditionally, statistical offices have been able to hide behind their publications. A statistical survey is responsible for the contents of the reports and publications that it publishes, but it takes no responsibility for how the user may possibly combine the data in the publication with data from other publications. This strategy will not be maintainable in the era of new technology.

This is a difficult problem, and there is no simple solution to it. However, it seems clear that statistical offices must activate themselves in the area of **standardized concepts and classifications**, an area whose importance is not always fully appreciated by those in charge of individual surveys, and sometimes not even by managers on higher levels. On the other hand this seems to be an undisputed area in the sense that most people outside a statistical office accepts the office's responsibility and welcomes its competence in this area.

Unified concepts and classifications is an excellent basis for combining data and putting them into models. However, there will always be cases where complete standardization is not possible, and it is important that a statistical office plays a constructive role also in such situations, even though it may not itself be responsible for the difficulties. For example, due to the different purposes of an administrative register and a statistical survey, it is inevitable that all definitions cannot be harmonized between them. However, experts in statistical offices should take as their responsibility to find ways around the problems, exploiting in a positive way the power of statistical method.

By means of these examples I have shown that there is a need for managerial action and control that is **global** in relation to the individual surveys. One may go one step further and say that the new problems and possibilities call for a **new survey concept**. Traditionally, a survey has been modelled as a serial flow of production steps, starting with survey design and data collection, and ending with tables, reports, and analyses. One effect of modern technology is that the ties between input and

output will be weakened, both physically, logically, and in time. The statistical end-products and typical usages of statistics will be based on combinations of input data from many different sources, and the data collected by one statistical survey will be used for many different purposes, by different users, and at different points of time.

Thus, if we look upon a statistical survey as a basic building block of statistical organizations and statistical information systems, it may be more adequate to think in terms of three different types of surveys:

* **input-oriented surveys**, collecting and editing the data, performing some routine tabulations and analyses, and preparing the data for future use by putting them, with their accompanying metadata, into common databases;

* **common databases**, taking care of data from different input-oriented surveys, and forming the basis for output-oriented surveys;

* **output-oriented surveys**, making use of existing data in common databases and other sources, inside and outside the statistical office.

**Latest R & D Reports (area ADB) published by Statistics Sweden:**

E-15 Using the RAPID Data Base Management System in Statistical Offices (Bo Sundgren)
E-16 Cell Supression in multi-dimensional frequency tables (Hans Block)
E-17 On Requirements for a Conceptual Schema Language (Björn Nilsson)
E-18 The Impact of Microcomputers on the Statistical Environment (Björn Nilsson)
-"- On the Usage of Microcomputers in Developing Countries (Björn Nilsson)
E-19 Conceptual Design of Data Bases and Information Systems (Bo Sundgren)
E-20 Stepwize formalization of information specifications by extending a simple object-oriented approach (Erik Malmborg)
E-21 Outline of an algebra of base operators for production statistics (Bo Sundgren)
E-22 How to satisfy a statistical agency's need for general survey processing systems (Bo Sundgren)
E-23 A session with the CONDUCTOR (Bo Sundgren)
E-24 Useroriented Systems Development at STATISTICS SWEDEN (Bo Sundgren, Birgitta Lagerlöf and Erik Malmborg)
E-25 On the semantics of aggregated data (Erik Malmborg)
E-26 Using the base operator approach for editing of statistical data (Bo Sundgren)
E-27 Towards improving communications between statistical evaluation of some communications options and details of STATCOM: A trial using 'com' computer conferencing and E-mail software (Jonathan Palmer)
E-28 The labour force survey in Zimbabwe - an illustration how the SCB model could be implemented in practice (Milan Sanovic)
E-29 An approach to the design of the time concept in the SCB model (Milan Sanovic)
E-30 The impact of the Development of EDP on Statistical Methodology and Survey Techniques (Lars Lyberg and Bo Sundgren)
E-31 SPORT-SORT - Sorting Algorithms and Sport Tournaments (Hans Block)
1988:3 Base Operators as a Tool for Systems Development (Bo Sundgren)
1988:4 Development of Systems Design for National Household Surveys - Report from a short-term mission to Harare, Zimbabwe, 12-28th January, 1988 (Birgitta Lagerlöf)
1988:11 Design of the User Interface for an Object-Oriented Statistical Data Base (Erik Malmborg)
1988:13 Education and training in the SAS System at the Central Statistical Office in Harare (Sten Bäcklund)

**Copies of these reports as well as previous reports still in stock may be ordered from Statistics Sweden, att. Ingvar Andersson, S-115 81 STOCKHOLM.**