STRATEGY FOR STATISTICAL COMPUTING FOR THE COUNTRIES OF THE CIS

Proceedings of the seminar for the heads of the statistical services of the Commonwealth of Independent States

Luxembourg, 4 to 8 July 1994

Sponsored by Eurostat/TACIS

Miscellaneous

Studies and research

OD

STATISTICAL DOCUMENT

Chapter 2.

STATISTICAL COMPUTING NEEDS

SPEAKER LARS THYGESEN, DIRECTOR, INFORMATION TECHNOLOGIES, DENMARK STATISTICS, DENMARK

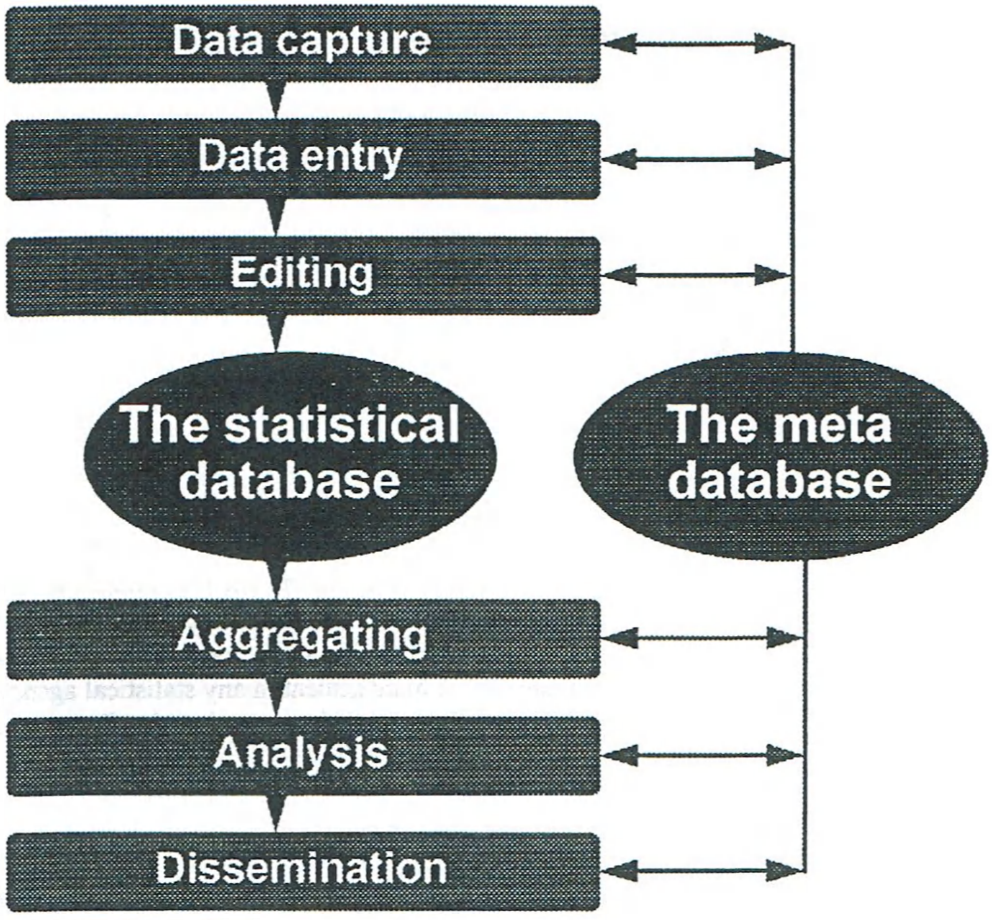
THE NEED FOR STATISTICAL COMPUTING, by L. Thygesen

1. INTRODUCTION

By its very nature, the production of statistics *is* data processing. Therefore, it is only natural that central statistical offices in most countries have had a leading position in the early development and practical use of IT methods in their countries - they have been the pioneers. This is no longer the case. But the efficiency of a central statistical office is to a large extent dependent on the way IT is used in the organisation. Consequently, IT must always be a key issue for the management in any statistical agency. It must not be left over to the specialists. The organisation of this high level seminar by Eurostat is therefore highly appropriate.

OUTLINE OF THE PAPER

Producing official statistics encompasses a range of well-known processes, all of which should be aided by IT. Indeed, the full gains of IT can be reached only when all of the processes are carried out in one integrated system. The diagram below gives a model of the processes of statistical computing. ***(See next page: schema).***



In this paper I shall focus especially on four of the work moments and try to illustrate the potential for benefits to be achieved. These work moments are: Data capture, editing, dissemination, and documentation. I shall briefly mention some of the promising tools available. Finally, I shall discuss the importance of integration, and some other strategic issues.

THE TASK OF THE CENTRAL STATISTICAL OFFICE

The platform for my discussion will be that of a central statistical office (CSO) of the same kind and with a similar position in society as that of my own office, Danmarks Statistik. I shall assume, that the production of statistics is fairly centralised and that the CSO is responsible for the co-ordination of statistics on society, covering a broad spectrum of topics: Population, business, industry, the environment and the national economy. The statistics are intended for many different users: The general public, central and local government, enterprise, research, and international organisations.

2. DATA CAPTURE & EDITING

The collection of basic data for statistics may be carried out in different ways and from diverse sources. An important distinction is between surveys and administrative registers.

SURVEYS

The taking of surveys is probably the most wide-spread method of data collection for statistics, involving some kind of questionnaire being presented to the units appearing in the statistics, either to all

of them or to a sample. Of course, taking a survey requires having a good frame of all units in the population, from which a sample may be drawn. The collection process has to be carried out and monitored, regardless of whether it takes place as a postal inquiry or as interview.

REGISTER BASED STATISTICS

But in many cases, the necessary information is already registered in some format within a public agency, e.g. for taxation purposes. In such cases, it may be expedient to use the administrative register as source of statistics. This presents many problems of methodology but it may result in efficient production of high quality statistics (Danmarks Statistik, 1994).

2.1 Surveys and the like

It is obvious that all stages of survey taking benefits from statistical computing: The establishment of a frame, the taking and administration of the sample, construction of the questionnaire, etc. If data are not captured in a machine-readable format, data entry has to be performed, transforming it into such a format before editing can take place.

An important problem of conventional survey-taking has been that a large number of separate processes have been involved. One survey form has to be found and examined by several staff members and at different points in the process. This leads to a waste of resources and - which is perhaps even more serious - a long and time-consuming production process. As we all know, timeliness is one of the most important quality parameters of many statistics.

EDI

One of the techniques that may be employed in order to avoid some of the work processes is electronic data interchange, EDI. This means that the respondents, notably the enterprises, are persuaded to deliver the data in electronic form in some pre-determined format. EDI techniques are being used increasingly for business exchange of information, and it seems possible to make the respondents happier and at the same time increase efficiency and speed in the statistical work process[1]. Such opportunities are not frequently met, and they should be taken.

A similar technique is being used in one of the most important new statistical censuses in Denmark, the wage statistics. These statistics have been reorganised in 1993 in co-operation with the Danish Employers' Union. A common format has been agreed upon, meeting the demands of the official wage statistics as well as those of the Employers' Union. The data asked for are in accordance with the data kept by almost every employer in their own wage systems. The data collection, comprising all workers as well as white-collar employees, is then carried out in electronic form, either via the Union, or for non-members by direct communication with the employer.

CATI AND CAPI

Another important technique of streamlining surveys is computer-aided interviewing, either by telephone (CATI) or by personal interviewing (CAPI). In this way data are entered in machine-readable form simultaneously with the interviewing. The interviewer is routed through the questionnaire, conditional on the answers. And validity control and editing can be carried out in the same process, entirely or in part.

[1] Lebaube (1992)

In Denmark, this technique is now being used in all our surveys among persons, using the Dutch Blaise system[2]. This has proven to be a big step forward in the sense that it has saved costs and speeded up the publishing of final results considerably.

THE ELECTRONIC QUESTIONNAIRE

In 1994, Danmarks Statistik has started using the Blaise system in a new way, i.e. as an electronic questionnaire that is sent to respondents on a diskette. The respondent goes through the questionnaire and fills it in on his own micro computer. When the diskette is returned to Danmarks Statistik, the respondent has already entered the data and has been confronted with inconsistencies and apparent errors. In this way we can benefit from the same advantages as with CATI. The respondents are municipalities or private enterprise units.

2.2 A register based statistical system

Administrative records kept by public agencies have always been an important source of statistics. Of course, public administrative registers have also been heavily computerised in recent years, making it possible to obtain data from this source in electronic form and avoid time consuming data entry. But under some circumstances the usefulness of registers can be increased far beyond that point.

THE PERSON NUMBER - A COMMON KEY

In Denmark, human statistics are largely based on the administrative registers, meaning that surveys are seen as a supplement. Danish registers are especially well suited for the compilation of statistics[3]. The Central Population Register, CPR, operated by the Ministry of the Interior, contains basic data used by every public authority, including a unique and unambiguous identifier, the Person Number. This makes it possible to link data from many different administrative registers. Danmarks Statistik makes use of the CPR, tax registers, wage registers, social benefits registers, a central register of buildings and dwellings, and many more. The Act on Danmarks Statistik grants access to these registers which are kept by many different authorities.

The use of this system means that all human statistics are coherent and that even the population and housing census is based solely on the registers. It is a system that is highly efficient and that reduces the burden of response. It may be considered to create such systems in other countries.

3. DISSEMINATION

I shall now jump to dissemination, skipping the job of actually compiling the statistics, i.e. producing the relevant tables and graphs, and analysing the results. This is not because this is an unimportant task or because statistical computing is not involved - on the contrary! It should also be stressed that the planning of the compilation of results must naturally be integrated with the planning of the other phases of statistics production. But as mentioned earlier, I have to restrict myself and focus on a few issues.

The purpose of the statistical activity is of course to *present information to the users*. Information in this context means aggregated data of some kind in a format that is understandable to the user, and allowing the user to extract answers to the questions that are relevant to him. In addition to the aggregated data, there must be some data explaining the contents of the data, cf. the following

[2]Schuerhoff (1993)

[3]Danmarks Statistik (1994)

paragraph, and there may be a smaller or greater element of analysis, i.e. text, graphs, estimates of parameters in some model, etc.

It is evident that the results must be made *available* to the user: He must have access, and preferably cheap and easy access in a format that makes it easy for him to process the information further. The access may be conditional on payment or not.

For many years, dissemination has largely taken the form of books, or other publications in print, and this is still the case even though things are now changing. For the production of books we should of course employ modern computational methods, using desk top publishing to integrate the different elements of the publication. The author of an article or a book should be allowed to make the product in a final form, without the interception of print-shop experts or editors, in order to make publication as quick and efficient as possible.

It is evident that electronic dissemination of statistics is becoming increasingly important. As users become more and more computerised, they want to have also the statistical information in a format, that allows them to easily transform the data, process it further, and integrate it with information they have from other sources. In many cases, they want to include statistics in their own information product for sale on the market. The increase in demand for statistics on electronic dissemination media has been somewhat slower than expected ten years ago, and consequently the development of these media has not gone far. But judging from the present trends, it seems fair to expect that electronic media will dominate the market for official statistics in five years time, as far as professional users are concerned. Therefore, it is important to start developing the services now

OFF-LINE DISSEMINATION

Electronic dissemination may take different forms. One form that has been quite common for many years, is the exchange of an off-line medium, like a tape or a diskette, containing one or a number of tables. In some countries (but not in Denmark) it has also been customary to disseminate media with *micro data,* i.e. files of information on individuals (persons or companies) without identifiers.

The electronic dissemination mentioned so far is often characterised by a rather low content of analysis, as it contains only figures. But there is nothing preventing us from including more analysis, e.g. text or graphs. In this way we get an electronic publication that may resemble more or less the traditional printed publications.

The electronic services should comprise an efficient and user friendly search tool, allowing the user to find the information he wants. The search tool may be interactive so that the user is guided step by step to the goal, and the information is created in accordance with the user's wishes.

ON-LINE SERVICES

An important distinction goes between on-line and off-line services. On-line service makes use of data transmission via the telematic networks, allowing for very quick transmission. On-line services should be designed for the time-critical users who must be prepared to pay something extra for the speed of the service.

DATA BANKS

Starting in the mid 1980's, Danmarks Statistik has built a powerful dissemination machine: Big *on-line data banks*, consisting of very large dissemination databases covering every area of statistics, coupled with an interactive system for retrieval and aggregating of statistical tables. The data banks are redundant in the sense that they contain only aggregated data that exist in a more detailed form elsewhere in the statistical registers. But they are extremely useful in two respects:

1. They allow users to have direct access to large amounts of detailed statistics and to get an immediate answer to almost any question within the information power of official statistics

2. They present a strong co-ordination instrument inside Danmarks Statistik, forcing the different branches of statistics to conform with one format and to use the same concepts

In this way, they allow the staff from one branch of statistics to use other branches of statistics with a minimum of effort.

I believe that this kind of dissemination machine is part of the answer to the statistical needs of to-morrow. In addition to on-line access, users may of course have off-line services based on the data banks. As a rather recent example of this, mention may be made of a CD-ROM[4] produced in co-operation with the other Nordic countries.

The present soft-ware foundation for the Danish data banks is AXIS, running on an MVS mainframe, developed by Statistics Sweden. AXIS contains the necessary functionality for an efficient on-line service but lacks the modern look of to-day's systems. For the CD-ROM and other off-line services, and for further processing of on-line extracts from the data bank, the DOS-based system PC-AXIS is used, also from Statistics Sweden[5].

INTERNATIONAL DATA EXCHANGE

In relation to electronic exchange of statistical data, the international perspective becomes particularly important. In the European Union, large amounts of statistics must flow between the member states and Eurostat, as well as between one member state and another. This calls for a co-ordinated telematic network that will ease the transmission of data. But it also calls for development of standard formats for exchange of statistical data, including the necessary documentation. Eurostat and a number of member states have been very active in the development of such an EDI standard, GESMES[6]. I believe that large on-line dissemination data banks, like the Danish, with a link to the GESMES standard will make statistics flow smoothly between our countries. PC-AXIS actually supports export of statistics in the GESMES format.

I believe that on-line dissemination of statistics will gather tremendous momentum in the next few years, as *the electronic super-highway* will bring transmission prices down and speed up. In order to ease access to statistical data banks for the busy costumer who can't be bothered to use powerful search tools, facilities for easy standard extracts based on the user's specifications ("data shooting") should be developed.

---

[4] *Statistics Across Borders* (1994), published by the statistical offices of Denmark, Finland, Norway, Sweden and Greenland in co-operation with Nordic Statistical Secretariat

[5] Nordbäck (1992)

[6] Lebaube (1992)

4. DOCUMENTATION

The description of the data in the statistical system, as well as the processes in the system, is a particularly important part of the system itself. This part is often called the *meta database*. In spite of the fact that the documentation or the meta database has been recognised for decades as something quite essential to the efficient functioning of the statistical system[7], the situation in many countries is far from being a happy one. This seems to be due to the fact that the top management of the CSO tends to concentrate its demand on the "hard statistics", rather than thinking also of the flexibility and usability of the overall statistical system. It is evident, that this situation has to be changed[8].

In order to have an efficient statistical service, it is necessary to have a computerised system of all "global" meta data, i.e. descriptions of all cleaned and finalised statistical data: Micro data (statistical registers) as well as macro data (tables, aggregates). The meta data must have a direct (technical) connection to the data they describe, but they must deal with the information contents of the data. They must allow the *user* (in-house user as well as end-user of statistics) to understand the statistical contents of the data. The meta data should be used in the stage where a user makes a request for data and needs to know the possibilities available, as well as in the stage where a user receives the statistical information in order to use it in his own environment and for his own purposes. It should also be used in the very processing of data from one form to another.

It is important that the meta database should be *one coherent system*. It is not sufficient that each branch of statistics is documented in its own formats and according to its own guidelines. It is evident that this is complicated unless the organisation of computing in the CSO is very centralised; for a number of other reasons, it is not advisable to centralise computing in this way, cf. paragraph 5.1 below.

The documentation should also contain some standardised form of declaration of data quality dimensions.

Creating a real meta database that will meet these demands is not an easy task. Many resource-consuming projects aiming at creating *the meta database* have failed, among them one in my own office. But work is being done in several countries. In Danmarks Statistik a project aiming at establishing a coherent system has just been started and is expected to show some results during 1995. However, the total database is not foreseen until the year 2000.

If this project (and other, similar projects) is to succeed it is absolutely necessary that the management supports it wholeheartedly, and year after year.

5. A FEW STRATEGIC REMARKS

It is impossible to discuss the needs for statistical computing without including a few remarks on the overall IT strategy of the CSO.

5.1 Organisation: Centralisation or decentralisation

There is a tradition in statistical services, including the Danish, of organising computing in a centralised way. Development and maintenance of the systems necessary for collecting, processing, and

[7]This can be seen from an great number of international statistical meetings in the UN, focussing on meta data

[8]Sundgren (1994)

disseminating the different statistics has been seen as a specialist job that could not be handled by the statistical branches. Consequently, there would have to be a central computing department.

In the centralised organisation, there is a strong professional tradition in the computing department. The concentration of skilled staff will create an environment that supports professionalism and helps the systems analysts and the programmers to keep up their skills.

The problem that has appeared in many CSOs is that the distance between the statisticians and the computing people tends to become too long. The computer specialists may find it difficult to understand the contents of the task they have to perform. The statisticians, on their side, have little understanding of the technical problems of computing. A lot of explaining must be done in order to ensure that the computer systems really solve what they are supposed to solve. This has in many cases lead to distrust between the two parties, and to inefficient work.

According to the author's experience, a large degree of decentralisation is advisable. The statistical branches should command the computer specialist resources necessary to develop their own systems.

5.2 Mainframe or network

Another aspect of centralisation or decentralisation is the choice between mainframe and network. I believe that there is a strong tradition for being mainframe based in our statistical offices. But micros in networks offer a lot of new opportunities.

The mainframe based systems have been very solid and error resistant. Also data protection can be kept at a maximum level. On the other hand, the micros offer user friendly and flexible tools that allow us to adapt quickly to user needs.

Therefore, there is no future without the micros. We must have them in networks for several reasons, among other things in order to allow statisticians to make use of powerful communications opportunities. Looking five years ahead, I see no future either without the mainframe which must be included in the same network.

5.3 The importance of internal standardisation

As soon as the centralised model of one big mainframe computer is abandoned, the problem of standardisation of computer systems becomes very important. The vendors offer many different tools and every user may develop his own preferences. This may inhibit job rotation, or make it necessary to re-develop a statistical sub-system every time a new person takes over responsibility. It also makes it difficult to support the soft-ware products and give an efficient training. In Danmarks Statistik standardisation is helped by the decision to run media-less micros in a network with a lot of common support. All micros boot from the central servers.

In addition to the specific statistical work moments, we also have to run an organisation, involving a number of administrative routines well-known also in other organisations: Budget control, administration of staff, handling of mail, archiving, etc.

It is important that all of these tasks or job moments, the statistical as well as the administrative, should be interlinked. This means that the computer systems should be able to communicate with one another. Otherwise, double work will occur, e.g. the same data having to be entered into machine readable form more than once.

5.4 Outsourcing or in-house computing

A discussion that is taking place in the CSOs of many countries, including my own, is whether computing should be outsourced or not. The argument is often heard that "we must join the outsourcing wave" that has apparently hit private enterprise. The justification for this outsourcing movement is that the efficient organisation should concentrate on its *kernel activity*, i.e. perform only the tasks where it has top expertise, and for a CSO, this means *statistics*. Other activities should be bought outside the organisation from vendors who are experts in just those activities.

I am absolutely convinced that this is a mistake when we talk about statistical computing. As I have explained earlier, I believe that computing is the very heart of official statistics. If we leave the statistical computing to others, we loose control of our most precious resource. This mistake might prove just as fatal as that of some sovereigns of medieval Italy who built their power on a hired army.

6. THE CSO AS PART OF THE TOTAL INFORMATION SYSTEM

It is important to keep in mind that the CSO should be an integrated part of the total information system. This means that the CSO is part of the national information system, interchanging information with respondents as well as with users of statistics - incidentally, the respondents are very often also users of statistics. Information should be interchanged and processed in a way that is expedient to the whole system. At the same time, the CSO should be part of the international information system, meaning that information has to be interchanged to an ever increasing extent also with international partners.

REFERENCES

Danmarks Statistik (1994): *Register based human statistics*. Copenhagen (in print)

Lebaube, P. (1992): *EDI and Statistics - A Challenge for Statisticians.* Eurostat, New technologies and techniques for statistics. Bonn.

Nordbäck, L.(1992): *The PC-AXIS Vision, the Liberation of Official Statistics*. Eurostat, New technologies and techniques for statistics. Bonn.

Schuerhoff, M. (1993): *Blaise as a statistical control center*. Bulletin of the ISI, Vol. LV, Firenze.

Sundgren, B. (1994): *Et verksgemensamt datalager vid SCB?* SCB, R&D Report, 1994:1. Stockholm