

Topic (iii) - Administration and policy of statistical data confidentiality

## **ACCESS TO CONFIDENTIAL DATA IN AN INTEGRATED STATISTICAL SYSTEM**

Submitted by Statistics Denmark<sup>1</sup>

### **Contributed paper**

#### **I. INTRODUCTION**

1. Technological developments in recent years has led to a situation where data on almost all aspects of the individual person are available from administrative registers. The content of the different registers on individuals, which make the Danish statistical system, reflects that. So a large number of data and especially a large number of data combinations made possible by means of the personal identification number form the basis of the statistical products, and the customers are quite aware of these possibilities. There has been and are still an increasing demand for statistics on various combinations of data. Each statistical register contains the data necessary to cover a certain field of statistics but the need for combinations goes beyond these limitations.

2. The demand for statistics on combinations of variables from different registers has led to a discussion on how to be prepared in the best way to meet these demands. If you have to process data each time combinations are requested, processes have to be repeated again and again which leads to a waste of resources and apparent unnecessary delays which are hardly comprehensible to the customers.

3. A technical solution is very easy to find. You can store all data in one or perhaps a few data bases but that solution does not sufficiently consider data confidentiality and some uncertainty may arise in the population if all data concerning a single individual are accessible to a number of staff members of the statistical office. Some people may feel it like "big brother is watching you" and that could easily harm the interests of the statistical office.

#### **II. THE ORGANISATION OF THE DANISH STATISTICAL SYSTEM**

4. The system dealing with social statistics in Statistics Denmark is organised in a number of traditional files. Each one of these contains variables which are relevant to a specific field of statistics. By organising the system in this way, it is easier to limit the number of data which are accessible to a single staff member.

5. As a part of the security measures adopted by Statistics Denmark the access to confidential data is restricted. Only a few staff members, 4 at the most, are authorised to access a file containing such data. So only a limited number of data are available to the accessing person.

---

<sup>1</sup> Prepared by Finn Spieker.

6. The Danish system has from the beginning been seen as a coherent system, where selected data from different registers could be linked by means of the personal identification number, but these identifications must be deleted from the file as soon as the link process has been carried through and the results accepted. Such files are still considered confidential but direct identification will only be possible for a short period following the link process. This strategy was decided at the beginning of the era of using administrative data for statistical purposes and it is still a basic part of the philosophy behind storage and utilisation of the data available for statistics.

### III. INTEGRATED REGISTERS

7. In recent years, there has been an increasing demand for data compositions across different registers. It has led to another strategy as far as the register system is concerned. Big registers crossing the traditional boundaries of the different statistics have been established. There are obvious reasons for this development. Storing data from different fields of statistics in one register means a higher level of readiness towards different purposes and a more efficient utilisation of the resources. Once certain data from different registers have been linked, you will not have to do that over and over again in order to meet different demands and if the output from a linking process is stored as a register, which means that personal identification numbers are stored as well, you will still have the possibility to link additional data from other registers or versions of registers.

8. However, the appearance of these integrated registers breaks with the principal model. Staff members authorised to access these registers will have the opportunity to access a very wide range of information on individuals. Most of the traditional statistical registers contain some data from other statistical registers but the recent development has been rather strongly in the direction of establishing a few real databases covering a significant part of the social statistics.

9. It is difficult to give a precise definition of these integrated registers but, besides the integration of data from various other registers, they are more or less characterised by being prepared for multipurpose use although they are named more narrowly. Some examples of the integrated registers are mentioned below<sup>2</sup>:

- **The Integrated Database for Labour Market Research (IDA)**

The main purpose of this register is to contribute to the statistical description of the Danish labour market and especially the changes from year to year. All relevant data concerning the labour force form part of the register and some special calculations have been made which are only available in this register. Besides these data, it contains supplementary individual information transferred from statistical registers of population, education, income, social benefits etc. The register is updated yearly and the first reference year is 1980. This register was the first of the kind and the use of it was intended to be very close related to labour market statistics, but it has turned out to be a data source for other statistical purposes because of the many supplementary data available for a relatively long period.

- **The Register for Preventive Medicine (FBR)**

This register has been established with special reference to health analyses in order to promote research in this field. It is possible to follow individual hospitalizations from 1976. Most of the data concerning health are collected from the National Patient Register and they are not available in other statistical registers but here again a lot of supplementary individual data such as family, demographic changes, education, employment, income, social benefits etc. have been transferred from other statistical registers.

---

<sup>2</sup> The names in short are related to the Danish names.

- **The Social Research Register (SFR)**

The purpose of this register is to gather selected individual data from different existing statistical register in order to rationalise the work of producing anonymised data sets (model data) to be analysed by the Social Research Institute with or without data from surveys carried out by the Institute itself linked at the individual level. From the beginning, the data in the register should only be kept for a period of seven years, but when the day for the first deletion of data arrived, strong arguments for not doing that were put forward by the costumer. So data will be kept for a longer period.

- **The Children Register**

Though the subject of the register as indicated is statistics on the conditions of children the content is more far reaching. The conditions of children depend on their parents. Therefore, data concerning parents have to be included in the register. Besides that, it will be part of the purpose to analyse how the circumstances during the younger years influence the later living conditions. So gradually the register will grow to cover the total population. Here again the data are collected from other statistical registers.

All the registers contain data on individuals and the PIN-codes are available if it is necessary to link data at the individual level.

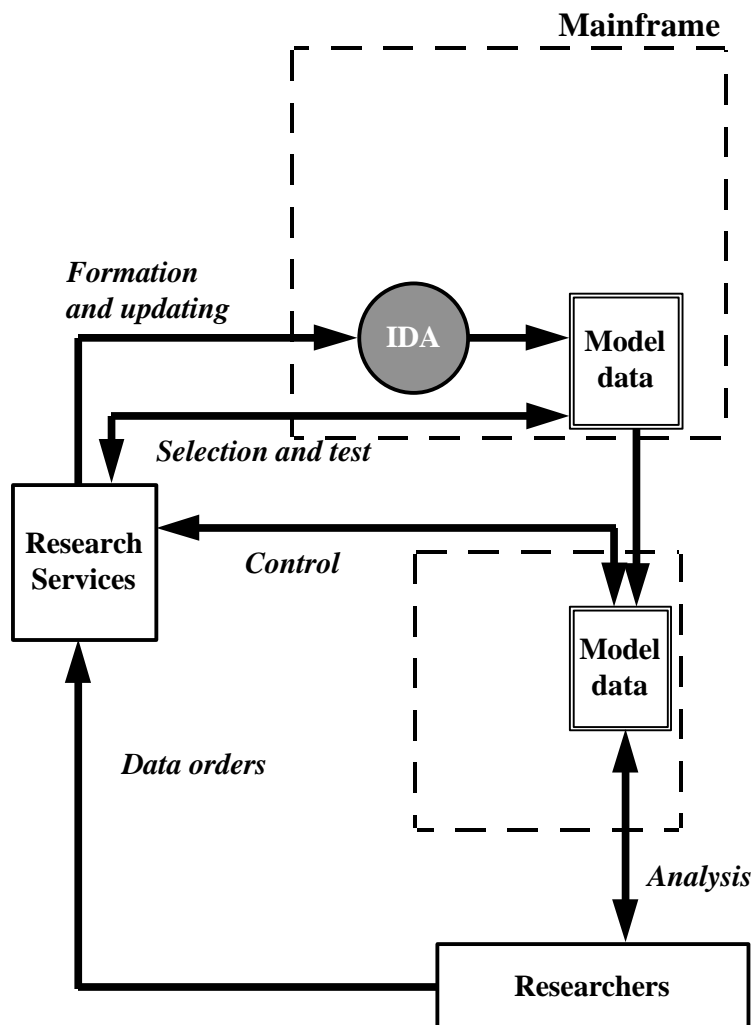
#### **IV. THE ORGANISATION OF AND USE OF THE INTEGRATED REGISTERS**

10. The integrated registers are organised in different ways. Some registers consists of one or more traditional files and others are organised as an Oracle database, and among the first mentioned you can distinguish between a single file model and multfile models depending on how data are stored. If all data have been linked once for all we are operating with the single file model. If data are kept separately corresponding more or less to the different sources we can talk about multfile organisation.

##### **The single file model**

11. Figure 1 shows an example of a single file model. All data are stored in one file (IDA) which form the basis for making the so-called model datasets which are available to researcher from work stations in Statistics Denmark on special conditions.

**Figure 1. The single file model of an integrated register**  
*The Integrated Database for Labour Market Research, IDA*

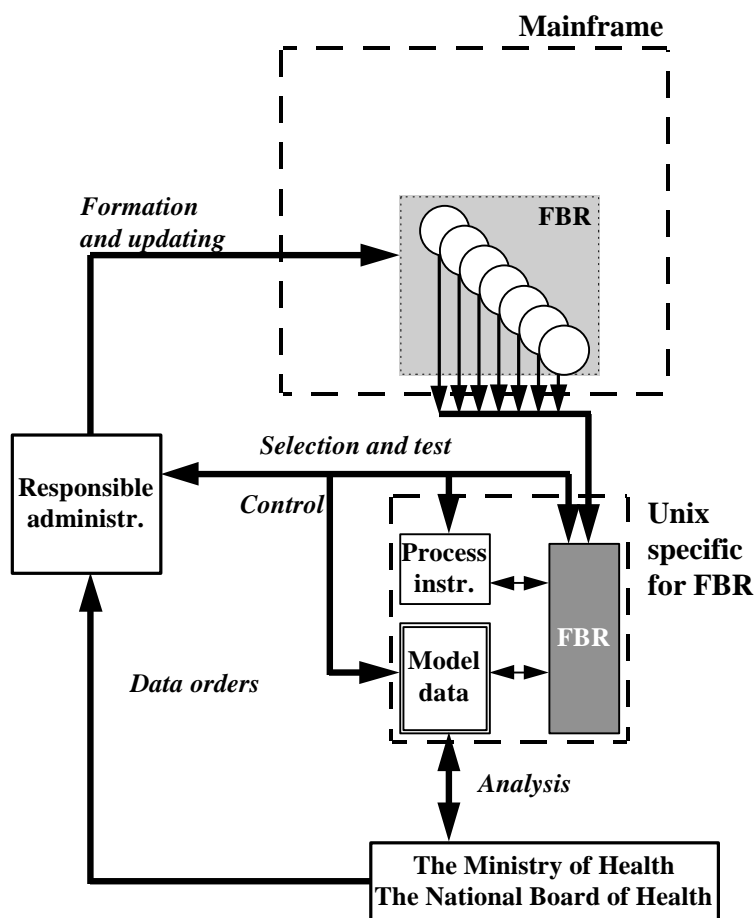


12. Research Services is responsible for the register and as such head of formation, updating and use. If a research project is approved and the data orders accepted by the research administration then the dataset will be made on the mainframe as an extract of the data required and the file transferred to the special computer at disposal for this kind of projects. The staff members who are authorised to access the register in order to update it or to form model datasets will have the opportunity to see many data on different subjects concerning identified individuals, and they do not have to link data or anything else to have the complete combination of data available. I has already been made.

### **The multfile model**

13. Using a multfile model is a little bit different. Figures 2 and 3 below illustrate how this model works. The difference between the way they are worked out are especially attached to the demands from various costumers.

**Figure 2. Multifile model 1 of an integrated register**  
*The Register for Preventive Medicine, FBR*

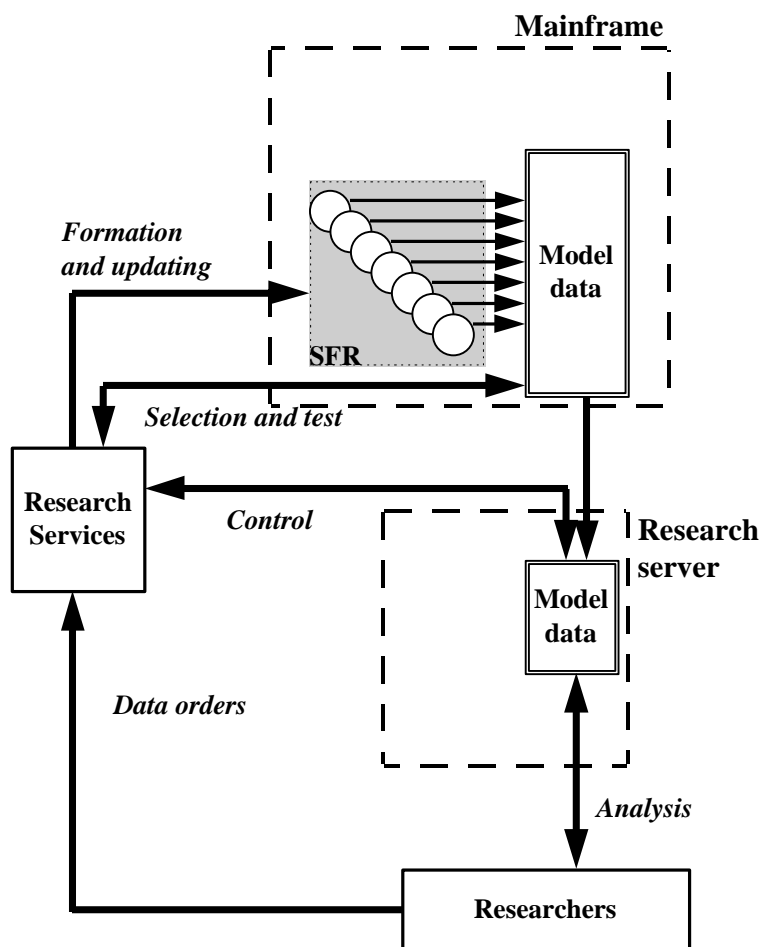


14. Both models store data in a number of files and for definite projects model datasets have to be formed by linking the relevant subfiles by means of the PIN-code. Though all data are available to the accessor as in the single file model there is the difference, that he will have to process data to read combinations beyond the limits of the single subfiles. Such a process will be logged and in this way a supervision will be more easy to carry out.

15. The difference between the two multifile models is the use of a Unix computer to serve especially the Ministry of Health and the National Board of Health. In order to enable rapid reactions to demands from the two institutions the establishment of the register is a result of co-operation between these institutions and Statistics Denmark. As part of the agreement model datasets should be available for authorised staff members of the two institutions from certain security areas by themselves. Though the register is placed on the Unix only staff members of Statistics Denmark have access to the register itself, and only these person can carry out the formation of model datasets.

16. Researchers from other institution can have model datasets at disposal but in this case the procedure will be as shown below in figure 3.

**Figure 3. Multifile model 2 of an integrated register**  
*The Social Research Register, SFR*

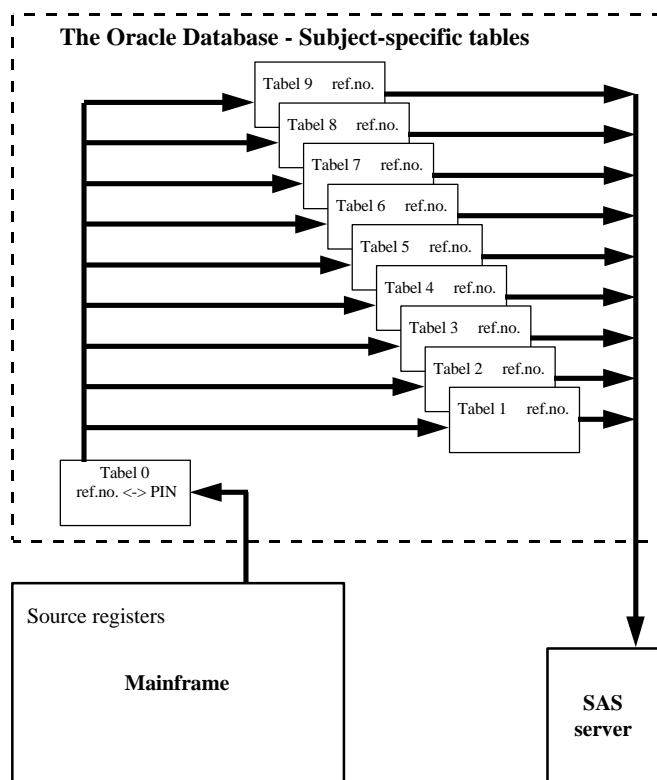
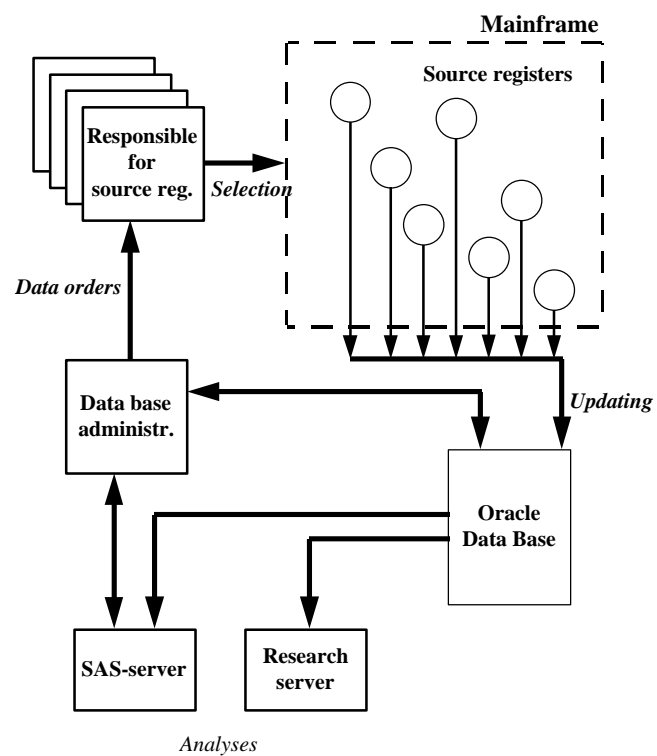


Multifile model 2 is somehow rather similar to the single file model as far as the final use is concerned and the conditions for having a model dataset for research purposes. But as described above, multifile data have to be processed to have combinations of data and that is the case in model 2 as well.

### The Oracle model

17. The Oracle model is the third one used in relation to integrated registers. Here the authorised staff members have direct access to all data included in the register. They can easily select all information on specific individuals. Even if a neutral serial number is used to identify the tables of the database a few person must necessarily have access to the table (table 0 in the lower part of figure 4) where the link between the serial number and the PIN-code is stored.

**Figure 4. Oracle model of an integrated register**  
*The Children Register*



In this model specified model datasets are formed like in the other models and they are transferred to a SAS-server, the research server or the common file server as SAS-data or Excel-data.

18. The use of Oracle has so far been limited. The technical advantages connected with such a system are striking but the philosophy behind the data security has been kept unchanged. So a conflict seems about to arise.

## **V. THE SECURITY EFFECTS**

19. The problems can be viewed from different angles. First, you must be aware of the security itself. The possibility to access individual data from different fields of statistics is not new. It has been the basis for the coherent statistical system. So the change is isolated to the way linked data are stored which means that a lot more different information on individuals is stored together in one way or another with the PIN-code attached to it. Still it is only a limited number of staff members who have access to these registers which helps to limit the risk. The possible reduction of the data security level caused by the use of integrated registers must be eliminated. It is the general opinion in Statistics Denmark, that these registers have not caused a perceptible change in the security level, but a little uncertainty calls for additional measures.

20. Secondly, it is important to focus on the public opinion. You must be sure that the security measures still are convincing. Even the smallest mistrust of the data security could have enormous consequences for the production of statistics in the future. It is a very important factor, and the existence of these integrated registers, maybe even bigger in the future, could cause some nervousness which must be avoided.

21. Thirdly, it has to be considered how new security measures will influence the data processing. Too many obstacles with limited importance for the data security would be incomprehensible to the users of statistics and will lead to dissatisfaction because of what is felt as unnecessary delay and in some cases incomplete products.

## **VI. SECURITY MEASURES**

22. Measures are to be imposed to solve the different problems mentioned above in a proper way, and the goal is to achieve a reasonable balance between the different considerations. Integrated registers and the activity attached to them are as all other statistical registers embraced by the Data Security Regulations for Statistics Denmark. Some of these common rules have been tightened up especially for these registers as follows:

- Only data which directly serves the purpose of the register must be included and the level of details must not be higher than what is necessary to carry out the expected analyses. The period of time must be restricted to what is relevant for the problem concerned.
- Data concerning crime or compulsory removal of children, which according to the rules must be stored only with encrypted PIN-codes, must not be included in an integrated register, while health data may be included if it is absolutely necessary and only after approval from the Data Security Committee of Statistics Denmark.
- The head of division responsible may authorise two staff members to access an integrated register. This number can be raised to three or four if the circumstances speaks for it and it is approved by the Data Security Committee of Statistics Denmark.
- Certain specific rules applies especially to integrated registers stored as Oracle database:
  - The official PIN-code must not be used as identifier between the different tables of the database.

The staff member responsible for the register must log all use of the table containing the links between the serial numbers which identify the tables and the PIN-codes. The log shall keep information on who have initiated a data process, when did it happen and what were the purpose.

23. Other supplementary measures have been discussed but it was felt that the ones mentioned sufficiently keep the data security unchanged at a very high and convincing level. Any security measures may here and there be felt as obstacles to the possibilities of carrying out statistical analyses but with these latest rules settled the different considerations are reasonably balanced.

## **VII. CONCLUSION**

24. From a technological point of view, the most effective way to handle the huge amount of data available for statistics from administrative registers would be to store all of them in one database accessible to all staff members who are authorised to access data in any field of social and demographic statistics. In this way you will escape from the redundancy which is very distinct in the present system.

25. In Statistics Denmark, the first step has been taken with the introduction of integrated registers and it has been followed by a project in progress which within a couple of years will lead to a complete metadatabase containing the documentation for all variables available in the statistical system. So it is very exciting to see what the next step will be. The very thought of this development frighten a little bit from a data security point of view, but you cannot walk against the wind. You will have to take advantage of the technological development and find solutions to the security problems.

26. It is quite difficult to foresee the real nature of the problems. Immediately it seems impossible to maintain the policy about access to confidential registers adopted by Statistics Denmark till now. Access to an Oracle database can be restricted to certain areas of the base, but it must be difficult to manage that if the restrictions have to be changed from time to time following the changes in the requirements for combination of data. A kind of temporary authorisation could be a solution but it seems not completely satisfactory. Hopefully an acceptable solution will be found in the near future.

### **Reference:**

Eurostat, Danmarks Statistik (1995): *Statistics On Persons In Denmark, A register-based statistical system.*