Working Paper No. 29
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)**

Joint ECE/Eurostat work session on statistical data confidentiality
(Luxembourg, 7-9 April 2003)

Topic (ii): New data release techniques

FROM ON-SITE TO REMOTE DATA ACCESS – THE REVOLUTION OF THE DANISH SYSTEM FOR ACCESS TO MICRO DATA

Contributed paper

Submitted by Statistics Denmark[1]

[1] Prepared by Otto Andersen (oan@dst.dk).

FROM ON-SITE TO REMOTE DATA ACCESS – THE REVOLUTION OF THE DANISH SYSTEM FOR ACCESS TO MICRO DATA

Otto Andersen, Head of Division, Research and Methods, Statistics Denmark

Summary

Statistics Denmark has altered its scheme for giving researchers access to de-identified micro data from on-site to remote access through the Internet. This is part of the general vision that Denmark should work hard to be one of the Worlds leading countries within registerbased research. Through the new scheme Danish researchers have experienced a breakthrough in the methods of access to micro data

1. From surveys to registerbased statistics

Denmark introduced the Person Number (the Personal Identification Number) in 1968 and it was used in a census for the first time at the Population and Housing Census in 1970. Accordingly, this became the first Danish register that uses the Person Number as an identification key. During the 1970s the first attempts were made to base the production of statistics on registers. In 1976 a register-based population census was conducted as a pilot project, but the registers were not sufficiently comprehensive and well-established until 1981, when a proper register-based population census was conducted containing most of the conventional population and housing census information.

Like in the other Nordic countries, the person and business registers in Denmark today cover a very substantial part of the production of statistics. The contents of the registers also cover many fields of research such as labour market research, sociology, epidemiology and business economics. The strength of the system is that the identification keys (person number, address, central business register number and property title number) render it possible to correlate the aggregated data both within a specific year and longitudinally across several years.

2. Increased interest in micro data

In the mid-1980s, Statistics Denmark experienced an emerging interest among various research environments and ministerial analysis divisions in applying micro data (individual data) for research and analysis purposes. One reason was that the development in computer technology made it technically possible to process large amounts of data according to advanced statistical models, such as multivariate models.

These environments put pressure on Statistics Denmark to disclose micro data; a request that Statistics Denmark was unable to grant because of the rules of confidentiality lay down by the Management and Board of Statistics Denmark. On the other hand, it was evident already at that time that not only were the registers of enormous importance to he production of statistics by Statistics Denmark, but their research potential was so great that it would be very valuable to actually utilise them for research purposes. Therefore, Statistics Denmark had to find a solution to the problem of access, which complied with the existing legislation on registers while taking into account Statistics Denmark's own confidentiality principles.

During 2001 negotiations between Statistics Denmark, the Ministry of Research and the Research environment resulted in a signing a contract on the establishment of a special unit (the Research Service Unit) in Statistics Denmark with the special duty to improve researchers access to micro data through a better infrastructure and to lower the costs of using the data. The budget for the Research Service Unit is 6 million D.kr. per year (approx. 800,000 Euro). Some of the money is used to upgrade the special Unix computers, cf. below.

3. Legislation

With the introduction of two acts on registers in 1979, Denmark saw the first statutory regulation concerning, inter alia, disclosure of micro data to researchers. As at 1 July 2000 these acts were replaced by the Act on Processing of Personal Data (lov om behandling af personoplysninger). The Act implements Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and the free movement of such data within the European Union. The former Act primarily governed registration and disclosure of data in registers, while the new Act applies to all forms of processing of personal data. The new term, "processing", covers all types of processing of personal data, including registration, storing, disclosure, merging, changes, deletion, etc.

Previously, the setting up of a register was subject to the so-called register provisions, involving a rather time-consuming and laborious process. These provisions have been abolished, and now the individual authority makes decisions in concrete cases on processing; for example, the authority decides issues of disclosure of data for scientific purposes based directly on the provisions of the Act on the lawfulness of such disclosure.

The new Act introduced a duty of notification to the Danish Data Protection Agency. The purpose is to enable the Agency to supervise the processing of sensitive information carried out.

Accordingly, a scientific project involving processing of sensitive personal data is subject to notification to and approval by the Danish Data Protection Agency before such processing can commence. This applies to all surveys, whether they are conducted by a public administration, individuals or enterprises. The Agency has laid down special provisions on security in connection with the processing of sensitive data.

All in all, the introduction of the Act on Processing of Personal Data has provided potentially more favourable conditions for register-based research in Denmark. In particular, public authorities' basis for disclosing administrative data for research purposes has been enhanced and simplified in terms of administration, as they no longer need to consult the Danish Data Protection Agency; personal data applied for statistical purposes may be disclosed and reused with the permission of the Agency; data from one private research project may be disclosed to another project; there is full access to filing of data in the State archives; both private individuals and public authorities may process data on Person Numbers for scientific or statistical purposes; furthermore, the Act now explicitly stipulates that the data subject's right of access to personal data shall not apply where data are processed solely for scientific purposes.

In addition to the Act on Processing of Personal Data, the Danish Public Administration Act (Forvaltningsloven) is of relevance. Under this Act, a public authority may impose a duty of non-disclosure on persons outside the public administration concerning the data disclosed. Statistics Denmark has applied this provision in connection with researchers' access to micro data under the scheme for the on-site arrangement for external researchers at Statistics Denmark (cf. below), although no disclosure in a formal sense is made. Data - even anonymised data - must be treated as confidential. Breach of the duty of non-disclosure is punishable by simple detention or imprisonment.

4. Confidentiality principles of Statistics Denmark

As it appears from the above, current legislation permits disclosure, to a wide extent, of personal data for scientific purposes. However, the authority in question ultimately decides whether disclosure may take place, meaning that the authority may take other issues into consideration even if the Danish Data Protection Agency has approved the disclosure of data.

That is what Statistics Denmark has decided to do. This decision has been made so that the individual citizen or enterprise can be certain that the data supplied directly or indirectly to Statistics Denmark do not fall into the hands of any unauthorised persons. In the opinion of Statistics Denmark the risk of

irreparable damage to the production of statistics outweighs the consideration for more or less convenient access to data by the individual researcher.

Thus, the fundamental principle is that data must not be disclosed where there is an imminent risk that an individual person or individual enterprise can be identified. This does not only apply to identified data, such as Person Numbers, but also to de-identified data, since such data are usually so detailed that identification can be made.

Since Statistics Denmark also considers it important that data can be applied for scientific purposes, special schemes for researchers have been set up.

5. Scheme for the on-site arrangement for external researchers at Statistics Denmark

Since its overriding principle is not to disclose individual data, Statistics Denmark set up a scheme in 1986 for the on-site arrangement for external researchers at Statistics Denmark. Under this scheme, researchers can get access to anonymised register data from a workstation at the premises of Statistics Denmark. Statistics Denmark creates the relevant datasets on the basis of the researcher's project description, the general principle being that the dataset should not be more comprehensive than necessary for carrying out the project (the "need to know" principle). The researcher signs an agreement which stipulates that data are confidential and that individual data must not be removed from the premises of Statistics Denmark.

6. Organisational framework

The scheme is administered centrally by the Research Service Unit as a part of the office of Research and Methods. The staff of this unit also create a substantial part of the interdisciplinary datasets and have a general (authorized) access to all relevant data in Statistics Denmark in order to reduce the administrative and bureaucratic work. The scheme requires close cooperation between Research Service Unit and the individual divisions. The advantage of such central organisation is that the individual researcher is fully aware of whom to negotiate with and who is responsible for the dataset supplied.

In 1996, Statistics Denmark opened a small branch in Århus, Jutland to grant researchers west of the Great Belt an opportunity to use the scheme on equal terms with researchers in Copenhagen. After having being funded by the Danish National Research Foundation (Danmarks Grundforskningsfond) Statistics Denmark has taken over the costs as a part of the above mentioned Research Service Unit.

7. Research databases

As the researchers almost invariably request datasets linking information from several individual registers in terms of both contents and time, the creation of specific datasets for a project often involves considerable work by Statistics Denmark and often considerable costs for the researcher.

To reduce the cost of datasets for research purposes and solve special data problems, Statistics Denmark has set up a number of research databases. These databases are hardly ever used in the actual production of statistics, but are first and foremost a kind of intermediate products for the benefit of the research process.

The most frequently applied research database is the Integrated Database for Labour Market Research (IDA). One reason for creating the database was to solve a difficult problem of definition: Identity of enterprises over time, a task that individual researchers were unable to handle for reasons of both time and funding. Nine to ten man-years were spent on the task, which was funded by the Danish Social Science Research Council (Statens Samfundsvidenskabelige Forskningsråd) and Statistics Denmark. Since the establishment of IDA in 1991, Statistics Denmark has handled the updating of the database against user charges.

Other research databases include the Demographic Database, the Fertility Database, the Prevention Register (health data), the Social Research Register, etc. As the names imply, the databases cover many specialist fields: economy, labour market research, social research, epidemiology, etc. The lasted development is the Prescription Database holding information on doctors' prescriptions of medicine sold by the pharmacies in Denmark.

A number of research institutes have paid for the creation of major research databases for the purpose of their own research.

8. Considerable growth

From the modest beginnings in 1986, the use of micro data has increased markedly under the scheme for the on-site arrangement for researchers at Statistics Denmark. In 1997, 71 researchers used the on-site arrangement, while in 2002 the figure had risen to more the 150.

9. Model and study datasets

Statistics Denmark has only to a very limited extent departed from the rule not to disclose micro data to researchers. To enable researchers to develop computer programs at their own workplaces, they have been granted an opportunity to borrow micro data, upon request, from very small populations (e.g., 1000 records). Only very few model datasets have been created in recent years.

However, Statistics Denmark has prepared some study datasets, so far based on the IDA database, for study programmes in economics/labour market policy and interdisciplinary data material for sociology studies. These datasets follow a few thousand persons over time according to a number of variables. Where possible, the data are scrambled so that the actual register data have been changed in ascending or descending order by a simple mathematical function. However, the fundamental characteristics of the data have been preserved. In this way, students get an opportunity to try out statistical models on realistic data.

Except for the above, Statistics Denmark has not applied scrambling procedures or special grouping techniques to the data that are made available to the researchers under the on-site arrangement. The data appear as in the basic registers.

10. The UNIX solution

Until 1996, researchers under the on-site arrangement were referred to making batch runs on Statistics Denmark's main frame. This meant that the only software available was SAS. Furthermore, most researchers were used to other platforms, such as UNIX, and therefore unfamiliar with the actual run and editing procedures.

In 1996, the Danish National Research Foundation funded the acquisition of a UNIX system, which has been used exclusively for projects under the on-site arrangement. The advantages were obvious: the researchers got access to known technology and the choice of software became more varied. Besides SAS, researchers now have access to SPSS, STATA, GAUSS, etc. Statistics Denmark has repeatedly upgraded the technical solution since 1996, partly by acquiring additional UNIX systems, partly by increasing the disc capacity. The latest and biggest upgrade was done January, 2003 as a result of the contact with the Ministry of Research on Research Service Unit

11. Remote access

In the autumn of 2000, the Director General of Statistics Denmark instructed a committee to examine whether to grant the users of Statistics Denmark's researcher schemes access to datasets from their own workplaces. The result of the committee work was a proposal to grant specially authorised research and analysis environments access to making batch runs on approved datasets of Statistics Denmark.

The Board of Governors of Statistics Denmark approved the scheme, which entered into force on 1 March 2001 following the completion of a pilot project.

A research or analysis environment can apply for an authorisation from Statistics Denmark. As at 15 March, 2003, 43 environments had been granted authorisation. The wording of the authorisation appears from Appendix 1.

Until now the remote access has not been granted for all datasets; particularly sensitive data (e.g., data on crime) has been excluded from the scheme and data on enterprises are assessed carefully to avoid any problems of confidentiality. It has been emphasised that the data consist of samples. If the researchers request access to total populations, the content of variables must be limited. The individual cases have been assessed by a steering group consisting of the Directors of Statistics Denmark. Researchers not granted remote access has been allowed instead to use the on-site scheme.

However, in December 2002 the Board of Governors of Statistics Denmark have accepted to a proposal to consider the on-site scheme and the remote access scheme as equivalent concerning data security and as a consequence of this decision all data sets which can be accessed from on-site can also be accessed from remote.

With this decision it has been very important to revise the rules for granting authorisation to micro data.

12. New rules for access to micro data

It is proposed to the Board of Governors (in it meeting on 2 April, 2003) that access to micro data can only be granted to researchers and analysts in authorised environments.

Authorisations can be granted to public research and analysts environments (e.g. in universities, sector research institutes, ministries etc) and to research organizations as a part of a charitable organization.

Within the private sector following user groups can be granted authorisation if they have a stable research or analyst's environment (with a responsible manager and with a group of researchers/analysts):

1. Nongovernmental organisations
2. Consultancy firms
3. Enterprises. However single enterprises can not have access to micro data with enterprise data

In order to grant an authorisation, Statistics Denmark will evaluate the proposed organization carefully and especially when it is an organization or firm within the private sector Statistics Denmark will look at credibility of the applicant (as ownership, educational standard among the staff and the research done for others).

Statistics Denmark will not grant authorization to single persons. Furthermore Media organizations are excluded from the scheme.

The "need to know" principle is still in force.

Researchers can have access to relevant business data after the "need to know" principle. Only very few business data are excluded from remote access. The whole question concerning these data are under evaluation.

13. Foreign researcher?

Only Danish research environments are granted authorisation as Statistics Denmark is not able effectively to enforce a contract abroad. Foreign researchers from well established research centres can have access

to Danish micro data from the on-site arrangement in Copenhagen or Århus. Visiting researchers can have remote access from a workplace in the Danish research institution during their stay in Denmark and under the Danish authorisation.

14. The remote access will take over

As a consequence of the decisions mentioned above the on-site arrangement will be closed down successively and the remote access will be the only route to micro data.

15. The technical solution

The technical solution is based on the use of the Internet conf. the flow chart at the end of this paper.

The relevant micro data are produced by the staff in Statistics Denmark and the de-identified micro data are transferred to the disk storage connected to the special Unix servers. These Unix servers are only used by researchers and are separated from the production network.

Communications via the Internet is encrypted by means of a so-called RSA SecurID card, a component that secures Internet communications against unauthorised access. In practice the researcher rents a password key (a token) from Statistics Denmark. The token ensures that only the authorised person obtains access to the computer system.

A farm of Citrix Servers ensures that the researchers from their own workplace can "see" the Unix environment in Statistics Denmark. All data processing is actually done in Statistics Denmark and data cannot be transferred from Statistics Denmark to the researcher's computer. The researcher can work with the data quite freely and can make new datasets from the original data sets. The limit is of course the amount of disk space. Statistics Denmark has just increased the total amount of disk space considerably.

All results from the researchers computer work can be stored in a special file and such printouts are sent to the researchers by e-mail. This is a continuous process (every five minutes) and has shown to be quite effective. The advantage to Statistics Denmark is that all e-mails are logged at Statistics Denmark and checked by the Research Service Unit. If the unit find printouts with too detailed data, contact is taken to the researcher in order to agree on details of the level of output. No severe violation of the rules, establish in the authorisation formula, has taken place.

.

Appendix 1

Statistics Denmark

AUTHORISATION

Statistics Denmark hereby grants

[Institution] represented by [Chief Researcher]

Authorisation for

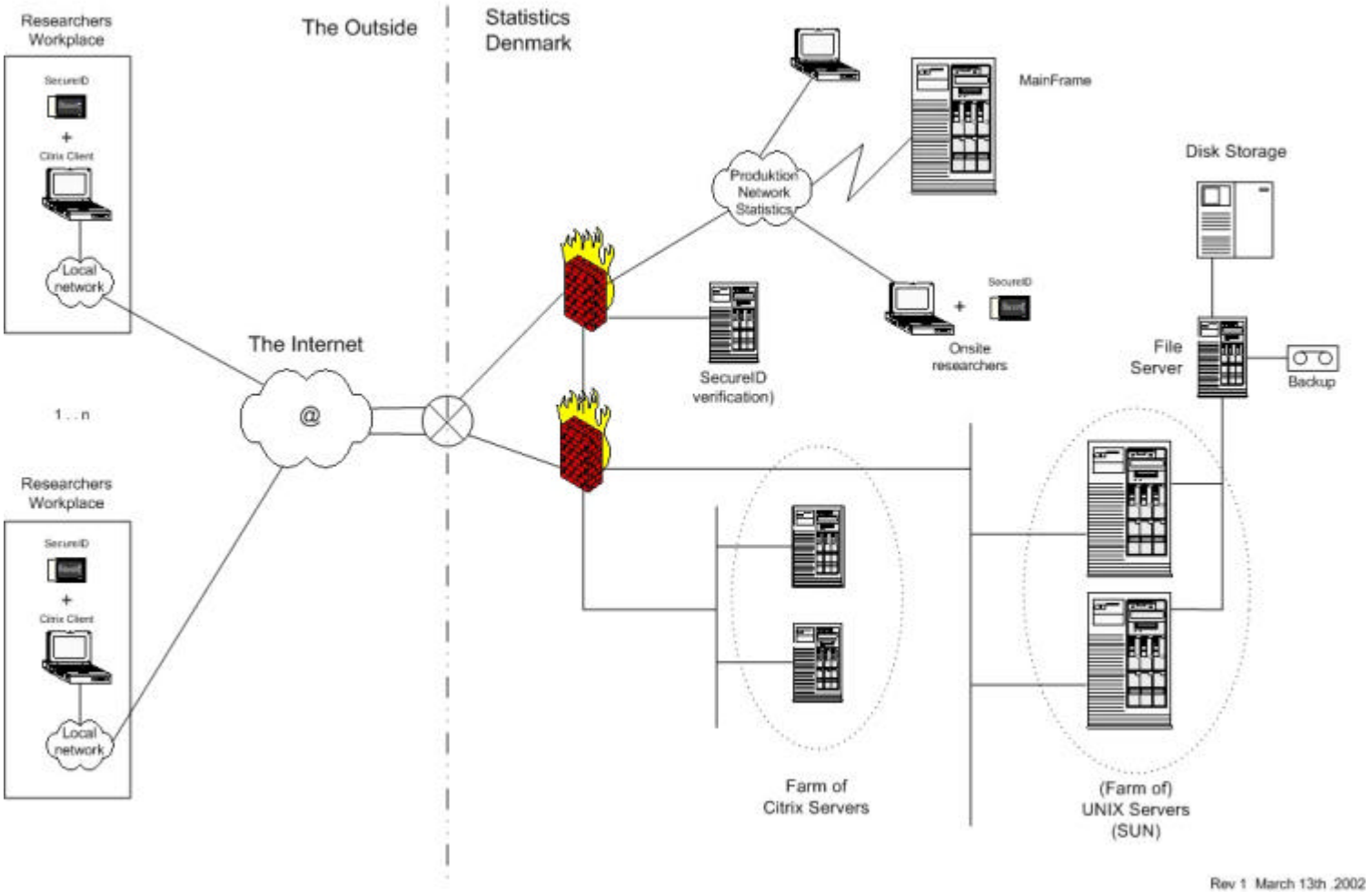
Remote electronic access to selected datasets at Statistics Denmark

Remote Access via the Internet is subject to the following terms:

1. A project description must be submitted, which states the project objectives and renders it possible to select the data required for successful project execution.
2. Based on the project and data description, Statistics Denmark decides whether external electronic access to data can be granted for the specified project. If the authorisation is not granted, the researcher is referred to use the ordinary scheme for the on-site arrangement for external researchers at Statistics Denmark.
3. The researcher to whom external electronic access is granted shall sign a special agreement with Statistics Denmark, cf. appendix.
4. All datasets are confidential, cf. §27(3) of the Danish Public Administration Act and §152 of the Danish Criminal Code.
5. The researcher obtains access to make batch runs on Statistics Denmark's special researcher machines (UNIX system) from one or more PCs specially assigned for that purpose in the research/analysis environment. Access is denied for batch runs from remote PCs, PCs at home or PCs which cannot be properly supervised.
6. Only the client software assigned by Statistics Denmark may be applied in connection with the RSASecurID card provided. A PC connected to Statistics Denmark may not be made available to unauthorised persons, and when the user leaves the PC, the PC must be either shut down or disconnected, i.e., protected from any unauthorised use.
7. The password of the individual researcher is personal and strictly confidential.
8. The researcher may not, directly or indirectly, download the dataset or any datasets derived there from. All transfers of output for printing or further statistical processing (in spreadsheets or similar) must be executed in accordance with the guidelines and methods laid down by Statistics Denmark. Statistics Denmark will create a log file of such authorised transfers. Furthermore, individual records may not be printed, and all output must be aggregated to an extent that eliminates any risk of direct or indirect identification of persons or enterprises. The researcher may not attempt to make such identification.
9. Statistics Denmark shall be entitled at unannounced visits to check that the rules of this agreement are observed.
10. The person signing this agreement on behalf of the research/analysis environment shall ensure that publications by the environment do not contain any information that may identify individual persons or individual enterprises.
11. The person signing this agreement on behalf of the research/analysis environment undertakes personally to supervise or to appoint a person to supervise that the provisions of this agreement are observed.
12. In case of breach of the provisions of this agreement, the researcher in breach will be excluded from using any researcher schemes of Statistics Denmark permanently or for a period of not less then three years. Furthermore, in the case of breach hereof, this authorisation will be withdrawn for a period.

This agreement, which is signed in two copies, enters into force on [date] and may be terminated by either party at three months' notice.

Remote Access to Statistics Denmark. January 2003. Principles of Operation



Rev 1 March 13th 2002
ABJ/-