Working Paper
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Helsinki, Finland, 5 to 7 October 2015)

Topic (iv): Access to Statistical Data for Scientific Purposes

Creating a National Remote Access System for Register-based Research

Marianne Johnson[*]

[*] Researcher Services, Department of Standards and Methods, Statistics Finland, Helsinki, Finland, marianne.johnson@stat.fi

Abstract: National registers, kept by a wide range of register keepers, are a rich source for data in Finland. Statistics Finland obtains much of its data for producing statistics from these registers. Statistics Finland is, together with the National Archives of Finland, working within the infrastructure project "Finnish Microdata Access Services (FMAS)" on improving the possibilities to use the microdata gathered in these registers also for research purposes. Maintaining the confidentiality of the data is of course of highest priority. Statistics Finland has provided researchers with microdata through its remote access system since 2010. It is now, as part of the FMAS project, opening the remote access system to be used by other register keepers in order for them to provide their microdata to researchers through it. At the same time the technical maintenance of the remote access system has been outsourced to the Finnish It-center for computing. One of the challenges when linking data from several register keepers within the remote access system is to develop a process for each register keeper to replace the national PIN:s with project specific pseudo codes. .

1 Register data for research and statistical purposes

Official data are often in the format of electronic registers in Finland. A large number of registers are held on the national level so that they cover the whole population, all enterprises or other units of the society. These national registers, kept by several governmental agencies such as the National Institution of Health and Welfare, the Tax Authorities, the Social Security Institute etc., are rich sources of data for scientific research. The value of these microdata as research data sources is further increased by the possibility to link them together by identity numbers (PIN, Business identification number). National comprehensive registers, especially personal data registers which all use the same identity numbers, are a unique richness of the Nordic countries.

However, the available data could be utilized much more effectively in scientific research in Finland, if the register keepers could simplify the access to these data. The confidentiality of the personal and business data puts, however, special challenges on simplified delivery systems.

For statistical purposes Statistics Finland has the right to gather data from the different administrative registers. Much register-based microdata is thus available through Statistics Finland and Statistics Finland has developed ways to serve researchers and maintain confidentiality at a high level. Statistics Finland developed a remote access system in 2010 following the development done in Denmark and Holland. At the same time Statistics Finland centralized its researcher services and declared as a strategic goal to provide good services for researchers.

2 Obtaining data by Remote Access from Statistics Finland

Up to 2010 researchers wanting to use microdata on businesses and enterprises had to travel to Statistics Finland and analyse the data at the Research laboratory. Researchers interested in microdata on individuals could obtain an anonymized (both direct and indirect identification rendered impossible) sample of the data. Researchers were not happy with neither the geographical inequality of the Research laboratory nor with the "spoiling" of the data before it could be released. In 2013 the Finnish Statistics Act was amended: it no longer is necessary to anonymize data before releasing it to researchers, as statistical authorities may give permission for research purposes to such confidential data from which the statistical unit can be indirectly identified. Direct identifiers must still be omitted from the data. As the data researchers now are allowed to handle no longer can be classified as safe, more emphasis is made on making sure the research environment is safe. Statistics Finland continues to release anonymised data (small sample, top coded, coarse classifications …) on CD's and USB flash drives but with data where there is a risk of indirect identification of the units only use over the Remote Access system or at the Research laboratory (which is just a means for a researcher to access the Remote Access system if they can't otherwise) is permitted.

2.1 Data security measures with Remote Access

The underlying principle of the remote access system is that researchers are able to handle the confidential data, that they have obtained a permit for, through a secure connection to a server at Statistics Finland. Statistics Finland opens its firewall to connections from IP-addresses provided by the researchers in the contract. Connections cannot be made from other IP-addresses. The person who has been granted a user license obtains a user id and for each logon a new disposable password is sent to his/her mobile phone. The data can only be used for the purpose accepted in the decision. The researcher can see the micro data on the remote desktop and perform analyses with the help of the programs available, but cannot copy the data out from the system. Only keyboard and mouse signals coupled with screen image are transferred between the system and the researcher's computer. All data transfers into or out from the remote access system is handled by personnel at the Researcher services or the IT-support at Statistics Finland. The data or outputs are transferred by using an SFTP client (WinSCP) for secure file transfer between the local and remote

computers. The researcher can access the research project's files for the duration of the permit.

According to the obligation to maintain secrecy, the researcher must ensure that the research results contain no unit-level data or possibility of their disclosure. Researcher services apply a screening process of research results, which ensures the implementation of data protection in the print-outs produced by the researcher from the data. All output is screened and after screening sent to the researcher by e-mail.

3 The Finnish Microdata Access Services -project

Statistics Finland, together with the National Archives, proposed a new unified service for register research, the Finnish Microdata Access Services (FMAS), for the update of the national roadmap for research infrastructures in May 2013. The proposal was approved and as a research infrastructure the project obtained infrastructure funding for development of the services for the year 2014.

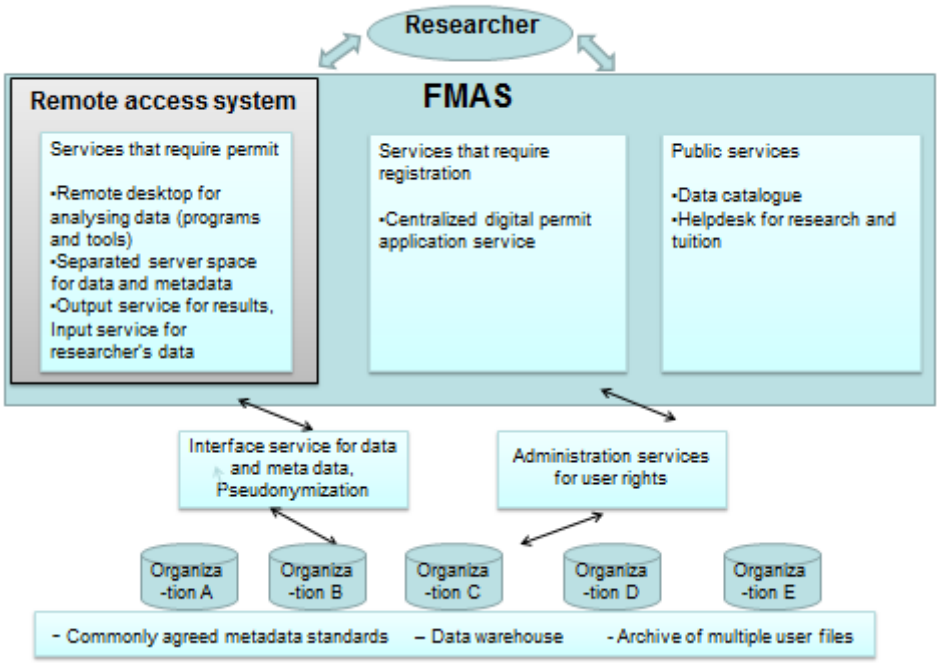

Fig 1. The Finnish Microdata Access Services.

3.1 Services provided by the research infrastructure

The Finnish Microdata Access Services, FMAS, is designed to facilitate the use of register and statistical data kept by different authorities. This service for researchers comprises four separate but interoperating services that provide a single channel for the entire research process, from planning to data analysis. It includes a metadata catalogue that can be used to find information on available data repositories, an electronic permit application service for obtaining permissions from various public authorities, a remote access system for the combination and analysis of various

licensed data, and an information and support service providing assistance and advice on all issues related to register research. The FMAS markedly simplifies finding, obtaining and using public administration data. It also improves the protection of personal data throughout the lifecycle of register-based research data.

4 Recent developments

4.1 Outsourcing of the remote access system

The remote access system has been developed and maintained by the IT-division of Statistics Finland. The directors of Statistics Finland stated early on an aim to find a more suitable organization for the technical management of the system. There have occasionally been some technical problems with providing remote access to a few researchers and the task of identifying the problems and providing solutions has at times been quite demanding on the it-personnel at Statistics Finland. This is obviously not the core competence area of Statistics Finland it-personnel.

The IT Center for Science (CSC) is a non-profit, state-owned company administered by the Ministry of Education and Culture. CSC maintains and develops the state-owned centralised IT infrastructure and uses it to provide nationwide IT services for research, libraries, archives, museums and culture as well as information, education and research management. CSC also has the task of promoting the operational framework of Finnish research, education, culture and administration. Presently, several services for researchers are being developed by CSC through a project on Opens Science and Research in Finland. This is an initiative funded by the Ministry of Education and Culture in order to create top conditions for research in Finland. These services include a data catalogue and storage space for research data files.

In 2010 CSC worked together with the Information Centre on Register Research on a proposal for a federated remote access system for providing researchers, in a secure way, with data from different register keepers. A working group appointed by the Ministry of Education and Culture stated that such a system is needed but that there should not be two separate systems developed in Finland. It was thus quite natural that Statistics Finland decided to outsource the technical maintenance of the system developed at Statistics Finland to CSC and together with CSC work on developing the system to become a national system. One of the first steps was to name the remote access system maintained by CSC the FInnish ONline Access -system (FIONA).

From outsourcing the technical maintaining of the remote access system follows that the microdata provided for researchers will be stored in servers at CSC. Statistics Finland and CSC have signed an agreement stating the scope of the assignment and what rights CSC have to the data. The personnel at CSC working with maintaining FIONA have been asked to sign a confidentiality commitment to Statistics Finland. Statistics Finland continues as the administrator of the system.

4.2 From Statistics Finland's remote access system to the National FIONA remote access system

As stated earlier, Statistics Finland requires that all data files, where there is a risk for indirect identification of the data unit, can only be used by researchers over the remote access system. Most other register keepers in Finland permit researchers to handle data files according to the Personal Data Act and the Act on the Openness of Government Activities. For specific research purposes it is possible, according to these acts, for researchers to obtain and handle data even including identifiers. Data from the registers have been handed over to researchers on e.g. DVD's or USB flash drives.

It is common that a researcher wants to build a research dataset by combining data from other register keepers with data from Statistics Finland. The register keepers have sent their data to Statistics Finland where all identifiers have been changed into matching pseudo codes that have been generated for the data abstracted from data at Statistics Finland. The same pseudo codes are used in the data set prepared for the researcher from data at Statistics Finland. This has also been the procedure whenever a researcher wants to link his or her data to data from Statistics Finland. Data from the other register keepers, the researchers own data and data from Statistics Finland have then all been made available to the researcher over the remote access system.

Part of the FMAS-project has been to make the principles and use of the remote access system known to the different agencies keeping registers. The aim of the national remote access system is that agencies can provide their data for specific research project directly through an interface to the FIONA system. Thus the register keepers would not have to send data over to Statistics Finland for the changing of identifiers to pseudocodes and the uploading of the datasets into the remote access system. Another goal is that the FIONA system would be used by other register keepers when making their data available for researchers instead of physically handing data over to the researchers even in cases when data is not being linked with data at Statistics Finland. The register keepers see advantages of using the FIONA system especially when it comes to data security and managing the data but researchers do still have doubts about the functionality (the robustness and reliability as well as performance of the hardware) of the remote access system and are put off by the extra cost for using the system.

4.2.1 Common pseudocodes

One of the aims for the national remote access system is that no data including direct identifiers should be sent from an agency to another or to researchers. It is, however, crucial that some method of linking data should be provided. Statistics Finland uses, when constructing pseudocodes, an opens source SAS-macro with different seed codes for each project. A plan is for Statistics Finland to send out the project specific seed code each time an agency is providing data through the remote access system that should be linked to data provided by Statistics Finland. The so called common ready-made data sets that Statistics Finland has produced for researchers pose a problem. Statistics Finland stores these datasets on the FIONA system's disk server and open up access to them for researchers that have obtained a permit. Thus the data set is not copied separately into each project's folder within the FIONA system which saves

disk space on the server. If Statistics Finland were to produce a project specific pseudocodes each time the researcher wanted to add other agencies data to the ready-made datasets this would increase the copies of the often very large ready-made data sets. If again Statistics Finland were to send out the seed code used for producing the pseudocodes in the ready-made data sets, several other agencies would gain knowledge of the pseudo codes used in the data sets. One compromise would be to send out the seed code also for the ready-made files but change it yearly when updating the ready-made files.

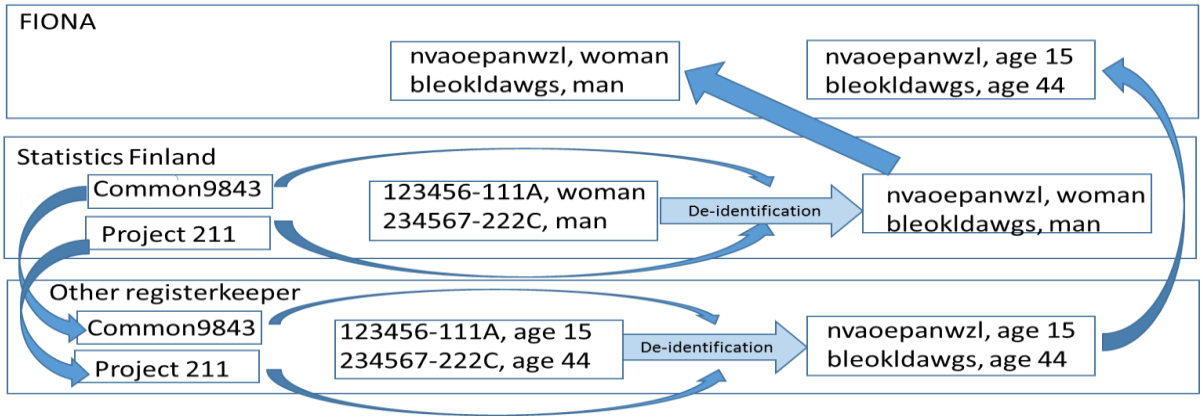

Fig 2. De-identified data uploaded to FIONA

4.3 Nordic co-operation

Statistics Finland has put out, together with the other Nordic Statistical Offices, a suggestion for making it possible for researchers to simultaneously use micro data from the different Nordic countries. The researcher should be able to obtain over the remote access system of either Denmark, Sweden or Finland data from all the Nordic countries. Data from the other countries would be sent to the statistical office of the country where the remote access system is situated, which would download the data set onto the remote access system and researchers from different countries could gain access to the data through the remote access system in order to analyse the data. There are step by step guides sketched out on how the statistical offices are to handle Nordic cases and how researchers should apply. Each of the different Nordic statistical offices have decided to trust the remote access system kept by another Nordic country as they are fairly similar in terms of technology and security. As Finland does output checking before sending out the results to researchers it was decided that this would also be the procedure in the Nordic co-operation scheme. As the persons in the different countries' registers differ from each other no linkable identifiers are needed in order to join the different countries data sets to each other.