United Nations ECE/CES/2013/33

Economic and Social Council

Distr.: General
10 May 2013

English only

Economic Commission for Europe

Conference of European Statisticians

Sixty-first plenary session
Geneva, 10-12 June 2013
Item 4 (a) of the provisional agenda
Drivers for micro-data access

Development and challenges of on-line micro-data usage

Note prepared by Statistics Finland

Summary

The paper describes development, current state and future challenges of Statistics Finland's services for researchers using micro-data. It mainly focuses on remote access services, but presents also a new microsimulation model for planning, monitoring and assessing the objectives and effects of personal taxation and social security legislation. Finally, the paper discusses the plans for creating a national system for accessing micro-data from governmental sources in different administrative sectors.

I. Background

A. Challenges

1. The increased demand for extensive confidential micro-data sets has created pressure for developing advanced modes of micro-data access for researchers. Researchers today need complex, easily linkable data in a technically advanced and flexible environment. The research community has emphasised the need for better, faster and cheaper services for researchers. Statistics Finland has received some feedback about the current practice of narrowing data contents and coverage due to data protection. High quality metadata and interactive information and support services are essential for research.

2. In 2011, Statistics Finland decided that *Improving Services for Researchers* is a strategic target for the office. All of the development work aims to service one purpose: the customer. Starting from the basics, customers should be informed about the data sources and variables, possibilities to access data, prices and timetable. There should be one contact point for researchers using micro-data instead of multiple contact persons in different statistical units. Furthermore, confidentiality of the personal and business data places a special challenge for providing researchers with easy access systems for using the relevant research data.

3. At the national level, a large number of registers covering the whole population of e.g., persons or enterprises can be linked together by common unit identifiers. However, this goldmine of information could be utilised much more effectively also in scientific research if the access to these data could be simplified. The challenge is to find common agreement on the financing and implementation of a national system serving this purpose.

B. Legislation

4. Data protection is a fundamental principle of official statistics, the objective of which is to ensure maintaining the trust of data suppliers and, thus, the availability of reliable basic data. The data protection rules that apply to Finnish official statistics are prescribed in the Statistics Act, the Personal Data Act and the EU Regulation on Community Statistics.

5. The Statistics Act regulates the compilation of statistics in Finland. Statistical production must primarily use data collected for other purposes in society. The vast majority of data are drawn from diverse registers. Only such data that cannot be obtained from elsewhere are collected from data suppliers. State authorities have a statutory obligation to supply their data for statistical production. Enterprises, municipal organisations and non-profit institutions are obliged to supply data on matters separately prescribed in the legislation.

6. The basic data for statistics are confidential and can only be released in a form from which individual units cannot be identified, and for scientific research or statistical surveys only. Exceptions to this include specific data items of the Business Register and the public data describing central and local government activities. With regard to personal data, data on age, gender, occupation, education and cause of death may exceptionally be provided with identification data for research and statistical purposes. An additional requirement is that the provision of data in an identifiable form is viewed as essential for the study. Confidential data may never be released for administrative decision-making or similar purposes.

7. According to the legislation, data on individual persons can be provided without identifiers within the scope defined by data protection regulations. A sample of data on a certain target group can be selected for researchers' needs using different information sources. At the moment, around 200 sets of micro-data, based mainly on register data relating to persons and housing, are supplied annually to researchers working outside of the statistical office. Data on the cause of death is the type of data most often requested by researchers. In addition to removing direct identifiers, the data must be made less detailed in order to prevent also indirect identification.

8. By contrast, until 2009 unit level data on enterprises and establishments could only be accessed at the premises of Statistics Finland under certain conditions of use. From the beginning of 2010, there has also been a possibility to use these data via a remote access system. Depending on the needs of the research project, access is usually granted to total enterprise data where direct unit identifiers have been removed.

9. Decisions on the provision of statistical data sets for research purposes are made by the Directors of the respective statistics departments or by the Director of the department of Standards and Methods where the research service unit for micro-data is located. In certain cases the Ethics Committee of Statistics Finland considers the requests for user licences. These cases include, for example, data requests from abroad, linking survey and register data in new ways as well as new practices in providing micro-data. The decisions on submitting micro-data abroad are made by the Director General. The terms set in the licence require that the data may only be used for the purpose indicated in the decision. The data shall also be treated as confidential, and they may not be released to others without authorisation from Statistics Finland.

10. The legislation regulating confidentiality protection is currently under revision. The goal is to harmonise the national legislation on statistics with the Regulation of the European Parliament and of the Council on European statistics. A working group appointed by the Ministry of Finance in 2010 has made a draft for the new Statistics Act and the law is expected to come into effect during 2013. The draft includes a proposal to promote wider utilisation of data collected for statistical purposes, especially in scientific research, but also in the education of future researchers and in general (public use files). The most notable change for researchers would be the possibility to obtain unit level data where individuals can be identified indirectly. As the data no longer would have to be anonymized by e.g. top coding, cruder classifications and sampling, this could increase the number of detailed variables and sample size in research data. For research services this would lead to a decreased amount of work on data protection when forming the data files. In order to provide a safer environment for usage of these detailed research data sets, remote access will become the principal mode of access.

II. Micro-data and remote access system in Statistics Finland

A. Micro-data for researchers

11. During the past two years, services for researchers have improved significantly at Statistics Finland. In June 2010, all research services providing access to micro-data were brought together in one department of Statistics Finland (in 2013 Standards and Methods) which enables combining social and business data more easily. The pooling of the resources and the newly enhanced processes have led to faster service and more satisfied customers. The service is financed partly by customers and partly from the budget of Statistics Finland. Furthermore, a new application for calculating prices has brought transparency and consistency to the pricing process.

12. Harmonised panel data sets make up the core of the micro-data available for researchers. The longest time series for establishments start in 1974 and for persons in 1970. Most of the enterprise-level annual micro-data sets are ready-made and available for research purposes easily. Different research projects may also order micro-data tailored for their purposes as a charged research service. In addition to data protection measures, this often includes linking data sets from various sources and building new variables. Tailor-made data sets thus take a longer time to produce and they also cost more. Tailor-made data are normally register data related to persons and housing or survey data related to living conditions etc. However, the aim is to increase the number of ready-made modules using person level data. Information currently available is described in more detail in Appendix 1.

13. All data sets are mutually linkable by encrypted unit identifiers (personal identification number, enterprise number, establishment number). They can also be linked to researchers' own data sets or data sets from other organisations or register authorities. In some cases, the researchers are able to link the relevant data themselves.

14. One of the most popular ready-made linked data sets is the Finnish Longitudinal Employer-Employee Data (FLEED). This unique database tracks and characterises the whole working-age population and their employees across two decades. In the personal data, identification codes for establishments and enterprises can be found based on the information of a person‘s employer at the end of each year. A protected sample of FLEED can be used through the remote access system.

15. Academic researchers are also very interested in obtaining individual earnings data. The Structure of Earnings Survey (SES) includes detailed information on the formation of employees' earnings linked to the background information on the employer. So far the SES data has been used e.g., in projects concerning earnings equality between men and women, segregation and earnings comparisons between labour market sectors.

B. Remote access system

16. The remote access system of Statistics Finland was piloted with the help of researchers in 2009 and put in operation in 2010. The aim was to improve data usage, data protection and regional equality among researchers. The main principles were adopted from Sweden, Denmark and the Netherlands, who have a lot of experience in developing remote access. In the system, the researchers use data on Statistics Finland's server via a secured internet connection (using SMS passcode) from their own workplace. Research organisations sign an organisation agreement for opening a remote connection to Statistics Finland and are responsible for ensuring their own users' compliance with the remote access rules. The contact person in each organisation is responsible for guiding the researchers on how to use the remote access system.

17. On the server the researchers can use a Windows desktop, where they have access to those metadata and data they have the right to use according to their user licence. At their disposal they have a separate working space, statistical programs (Stata, SPSS, R, SAS) and a folder for output. Servers are separated from the production network. The researcher cannot copy or transfer any data out of the system. The log files are saved and output is checked manually before it is sent to the researcher by e-mail. Currently, the system allows for around 16 simultaneous users.

18. A user licence is required in order to use statistical micro-data for research purposes. The applicant for a licence may be an official body, an institution or a person in charge of a study. Applications may also be filed by individual researchers. The applicant for a licence shall specify, in sufficient detail, both the purpose for which the statistical data are to be used, and the material requested from Statistics Finland, and any other statistical data that

will be used. A research plan and a pledge of secrecy shall be appended to the application. The researcher also signs an agreement with Statistics Finland stating the conditions of use, mode of delivery and prices. For remote access, the research organisation also has to sign an agreement on remote access including clarification of the data security practices in the organisation.

19. At the beginning of 2013, 12 institutes, 77 researchers and 43 projects had effective agreements for using the remote system. The feedback on the system has been mostly positive. Remote access has especially facilitated the use of business data and linked employer-employee data. In addition, joint projects of different institutes benefit from a common working space and efficient servers. However, the transition from using person level data in customer's own computer (e.g., data delivered on CD-roms) to using remote access is sometimes hard to accept by researchers. This is mainly because of the delay caused by the manual output checking procedure and the additional prices for using the remote access system. Furthermore, the possibility to use the system from a home computer or from abroad is not yet available.

C. Microsimulation model for income transfers and taxation

20. As a new approach to on-line data usage, Statistics Finland develops, maintains and distributes the static SISU microsimulation model for describing the personal taxation and social security systems of Finland. A new easy-to-use user interface was opened to users in April 2013.

21. The SISU model is a calculation tool intended for planning, monitoring and assessing the effects of personal taxation and social security legislation. Microsimulation models have been used for quite some time now in Finland in the drafting of legislation on social security benefits and income taxation. Microsimulation models can be used to calculate from unit-level data on a sample representing the whole population the overall effects of legislative amendments on different types of households as well as the whole population. The models are used to estimate tax revenues in the public sector, to examine the financial positions of individual persons and households, and to study income differentials and incentive effects.

22. The new model brings Finnish microsimulation to a completely new level in terms of usability and calculation accuracy. A new separate user interface was tailored for the model to enhance its usability. Calculation accuracy has improved along with the new larger basic data that better represent the population. The use of large register-based data is possible via remote access. The aim is to increase the use of the model in e.g. research and for assessing policy alternatives.

23. The model has been developed at Statistics Finland since 2011 in close co-operation with the Research Department of the Social Insurance Institution of Finland. Several experts from the Ministry of Finance, the Ministry of Social Affairs and Health, the National Institute for Health and Welfare, and the Government Institute for Economic Research contributed to the development work.

24. The SISU microsimulation model is composed of the main model that combines the whole income transfer system and of 12 sub-models that can also be used independently in simulation calculations. Each sub-model generally contains the taxes and benefits belonging to the same legislative collection. Either the service data of income distribution based on a smaller sample (a sample of around 27,000 persons) or a large register-based data set that better represent the population (a sample of some 800,000 persons) can be used in data simulation.

25. The SISU model can be used for scientific studies and statistical surveys within the framework of the Statistics Act. Restrictions apply for the data used in the model, but the actual model code is available upon request. In order to provide a secure and technically efficient environment for data usage, the model operates via remote access connection as the underlying data are very extensive micro-data and the person level data are visible to users. In addition, the model can be used with smaller service data directly from the researcher's workstation. Remote access requires the opening of a remote access connection, on which an agreement is made with the organisation. A separate agreement is made on remote access to the SISU model specifying the users of the model.

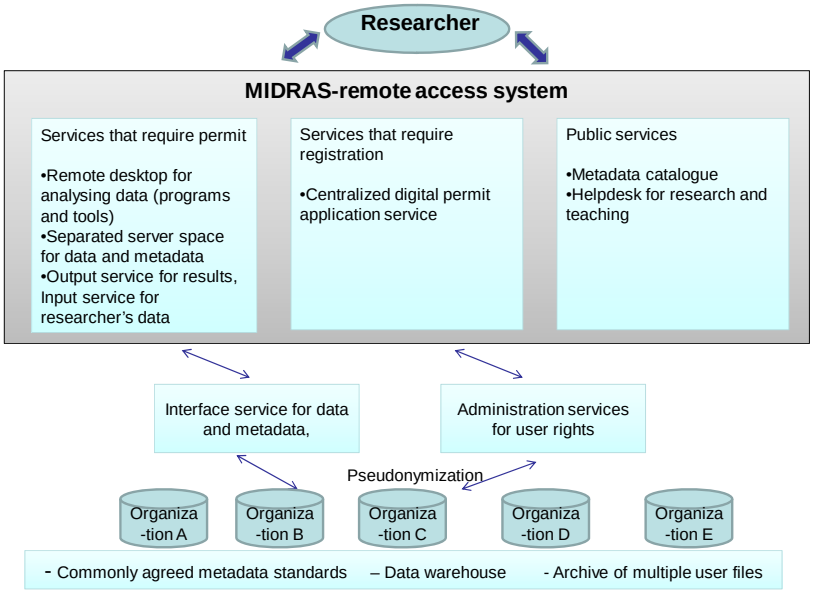
III. Finnish governmental co-operation in developing a national system for micro-data access

26. The rich interlinkable administrative register databases and survey data sets are goldmines for e.g. empirical economic, social policy and health analyses. Researchers often want to combine micro-data from several governmental agencies in different administrative sectors (e.g. Statistics Finland, Institution for Health and Welfare, Social Insurance Institution of Finland, Tax Authorities, Ministry of Employment and the Economy). However, the process for micro-data access is often complicated, expensive and time-consuming because of the diversity of legislation, application procedures and data distribution systems.

27. There have been several proposals for a national remote access system both from the research community and governmental agencies. On the basis of development work done in the last years, the National Archives, Statistics Finland and the IT Center for Science have submitted a proposal for a national project on Micro-data Access Services to the Ministry of Education and Culture.

28. The national system (illustrated in Figure 1.) would include services such as a joint metadata catalogue, a centralized digital research data permit application service, information and support service as well as a restricted remote desktop environment, including services and applications for processing and analyzing data sets. The Micro-data Access Service web-portal will be the main channel for obtaining information on possibilities for register-based research, guidelines for applying for access to data, the metadata catalogue and the digital application service. Thus, after submitting only one application per study the researchers would be able to combine high quality register data on, e.g. health, taxation, jobs and social reimbursements from different agencies using encrypted identifiers or pseudocodes (i.e. artificial codes). The updated metadata would be easily available in unified format allowing for searches in the catalogue for specific details on micro-data.

Figure 1
Micro-data access services



29. The project faces challenges both when it comes to the governance and identifying funds for the different parts of the micro-data access service. Also the co-operation and unification of practices between different register authorities will need to be developed. Statistics Finland is willing to take a central role in the governance and implementation of these services, but also other alternatives have been suggested. This new research infrastructure will bring research possibilities in Finland to a new level.

IV. Future challenges

30. The needs of the research community have to be taken into account in this national development process. In Finland the Ministry of Education and Culture is supporting the greater use of data through the Research Data Initiative. New national research infrastructures are being planned in order to facilitate the secondary usage of governmental data for research purposes.

31. During the last decades international mobility of researchers and the possibilities for studying or working abroad have increased considerably. Both Finnish and foreign researchers require access to micro-data also from abroad. Furthermore, we will need network solutions to allow the use of micro-data from international sources for research purposes, especially at the Nordic and the EU level.

32. Currently, the research services of Statistics Finland are exploring the experience of other countries with different kinds of technical systems, practices and legislation for micro data access. Accordingly, the work done by the project Data without Boundaries, funded by the EU Seventh Framework Programme (FP7), is closely monitored.

Appendix 1

Micro-data available for research use

1. Ready-made data files for research provide a diversified information basis for studying the characteristics and development of businesses and their employees in Finland. The following data and their combinations are available:

(a) Enterprise and establishment level statistical files of the Business Register;

(b) Enterprise Group Register;

(c) Financial statement data panel;

(d) Research and development (R&D) panel;

(e) Innovation data;

(f) ICT panels;

(g) Patent data;

(h) Business Aid Database;

(i) Commodity Statistics;

(j) Establishment based worker and job flow data;

(k) Establishment/enterprise based data on personnel characteristics and wages and salaries;

(l) Finnish Longitudinal Employer–Employee data;

(m) Data on the Structure of Earnings.

2. Register data related to individual persons and housing provide extensive scope for forming a variety of tailored data files for research. Tailored files can be formed from data on the following topics:

(a) Population (data on e.g. families, dwellings and housing conditions, buildings and holiday residences and employment);

(b) Education;

(c) Justice;

(d) Income and consumption;

(e) Wages, salaries and labour costs.

3. Service data files for researchers have been formed from the following important interview data describing the population and living conditions:

(a) Adult Education Survey;

(b) Time Use Survey;

(c) Households' Consumption;

(d) Households' Assets;

(e) Border Interview Survey;

(f) Use of Information and Communications Technology;

(g) Income Distribution Statistics;

(h) Quality of Work Life Survey;

(j) Labour Force Survey;

(k) Leisure Survey.