Business Structures Unit
Satu Nurmi

09.10.2006

Access to enterprise data at Statistics Finland: introducing the practices of the Research Laboratory

1. Background

The demand for micro data for research purposes has increased enormously in the past decades. Technological advances in data processing software and increased hardware capacity have enabled the use of extensive micro data sets in economic analysis. In addition, the linking of various data sets makes it possible to analyse complex economic phenomena using data from different sources. However, protection of confidentiality is a crucial issue when considering researchers' access to sensitive micro data.

Statistics Finland offers versatile business micro data for economic research purposes through the Research Laboratory of its Business Structures department. At the Research Laboratory enterprise and establishment data can be used at the workstations reserved for visiting researchers. The increased number of research projects in the 1990s created the need for establishing a supervised environment for accessing business micro data. As a result, development of data sets and practices for research purposes has been ongoing since 1996. The Research Laboratory became officially operational in 2001 and it has already become quite well-known among researchers in economics.

2. Legislation

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

The compilation of Finnish statistics is regulated by the Statistics Act. Data collected in other contexts must be primarily exploited for statistics. 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 data from the information in their possession. Enterprises, municipal organisations and non-profit institutions are obliged to supply data on matters separately prescribed in law.

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 are the data in the Business Register and the public data describing central and local government activities. With regard to personal data, data on age, gender, occupation and education may exceptionally be released with identification data for research and statistical purposes. An additional requirement is that the release of the data in identifiable form is viewed as essential with regard to the study. Con-

fidential data may never be released for administrative decision-making or similar purposes.

According to the legislation, the release of data on individual persons is possible within the scope of data protection regulations. A sample can be selected from the data on a certain target group for researchers' needs using different information sources. At the moment, around 200 sets of micro data are released annually to outside researchers based mainly on register data relating to persons and housing. However, in addition to removing identification data, the data may be made rougher or combined in order to prevent identification.

By contrast, data other than those publicly available on enterprises and establishments may only be accessed at the premises of Statistics Finland under certain conditions of use. Access is usually granted to anonymised total enterprise data according to the needs of the research project.

3. Access rules

3.1 Applying for a licence

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. In cases where the applicant is an official body or an institution, authorisation is granted to a specific, named person or persons.

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 shall be appended to the application wherever possible. Authorisation to use statistical data is usually granted for a given period. For this reason, the applicant shall specify the estimated duration of the licence required for the purpose intended. However, for justified reasons extension may be granted for the licence period.

When considering the granting of a licence to use basic data, Statistics Finland first determines whether the data can be processed at Statistics Finland to obtain the statistics requested by the applicant. However, in academic research projects using micro data this is seldom possible. In considering the application, account is taken of the applicant's possibility to obtain reliable results by using the required material. This includes both checking the suitability of the available data to the research problem and checking the background of the researcher. In addition, attention is paid to the value of the research to society and to the further development of research data.

Decisions on the release of statistical data sets for research purposes are made by the Directors of the respective statistics departments. In exceptional cases user licences may be considered in Statistics Finland’s Ethics Committee. These cases include, for example, releasing data abroad, linking survey and register data in new ways or new practices in releasing data. The decisions on releasing data abroad are made by the Director General. The data

may only be used for the purpose indicated in the decision. The data shall be treated as confidential, and they may not be relinquished to others without authorisation from Statistics Finland.

3.2 Conditions of use

In the case of enterprise and establishment data, a licence to use statistical data applies only within the premises of Statistics Finland. A mutual contract is signed in order to agree on the prices and conditions of use when renting a researcher workstation at the Research Laboratory.

Identification data shall be removed from the material for which a licence has been granted. For example, names and addresses of enterprises are removed from the data. In addition, unit identifiers are encrypted. If the material in the possession of Statistics Finland is to be combined with other materials, for example, the researcher's own data sets, this combining must take place within Statistics Finland. Statistics Finland shall remove all identification data from the combined material.

According to the contract, the researcher pledges to only publish the research results in a format from which information concerning an individual enterprise cannot be identified. To ensure this, the research results must be presented to the Research Laboratory prior to their publication. In addition, the publication should be sent to the staff. The researcher also pledges not to reveal or use for his private gain the information prescribed in law as confidential acquired by him in connection with the research. The obligation to remain silent shall also apply to the data processing programs and instructions related to the production of statistics when the disclosure of such information could endanger data security. The obligation to remain silent shall remain in force even after the termination of the agreement.

The researcher is not allowed to bring own diskettes or other saving devices to the Research Laboratory. Instead, all the files and output go through Statistics Finland staff. Internet or e-mail may not be used at the Research Laboratory. Output is manually checked according to the disclosure control principles before sending to the researcher. As a rule of thumb, three observations per group are required. In addition, other checks are made depending on the data source and the type of output produced.

4. Functioning of the Research Laboratory

The Research Laboratory is primarily meant for carrying out extensive academic research projects. There have been around twenty research projects per year at the Research Laboratory during the past few years. Most of the customers come from Finnish research institutes and universities. The projects are funded through different sources, for example the Academy of Finland or the Finnish Funding Agency for Technology and Innovation. The basic setting up cost of a project is €883. In addition, there is a rent of €18/hour for working at the Research Laboratory. Research services can be bought for €97/hour.

The central role of the Research Laboratory is to support the staff especially at the beginning of each project. Guidance is given particularly on the capabilities and restrictions of the data and their use. Some support is also given in respect of statistical programs and economic analysis. The projects may order micro level data tailored for their purposes as a charged research service. This often includes linking data sets from various sources and building new variables.

On the one hand, close co-operation with the researchers increases the opportunity to monitor their work and, on the other hand, the possibility to receive very valuable feedback for statistics production. Research use may bring out new ideas for gathering important data or enhancing, for example, the time series properties of the data. Analyses may also reveal deficiencies or even flaws in the data or documentation. Joint research projects with universities or research institutes have also increased expertise in the operation of the Research Laboratory. The Laboratory has also used outside experts as consultants in developing its data sets.

The research projects have studied various topical and socially important research questions. The topics range from studying productivity, innovative activity and other factors of business performance to labour market dynamics and firm demography. In recent years, the focus of interest has also shifted to the effects of globalisation and information and communication technology.

5. Research Laboratory data

Statistics Finland's enterprise data based on both register data and surveys offer a rich information basis for studying various features and development of Finnish businesses. A more detailed list of the data sets is included in Appendix 1.

The Research Laboratory's enterprise and establishment data contain information about the unit under examination concerning its:

- characteristics (e.g. industry, location, legal form, ownership)
- activity (e.g. profitability, indebtedness, production, use of inputs, investments, exports, R&D expenditure)
- average characteristics of staff by establishment (e.g. monthly pay, education, age, gender distribution, marital status, ownership of dwelling)
- heterogeneity of staff (e.g. distributions of pay, age and level of education)
- staff mobility (worker flows in and out, divided by source and target)

In addition, a combined employer-employee data set is also available containing varied information about the characteristics and work histories of enterprises' employees. In the personal data, identification codes for establishments and enterprises can be found basing on the information of a person‘s employer at the end of each year.

The individual-based information, based mostly on Employment Statistics, can only be accessed by the personnel of Statistics Finland. However, outside researchers may be granted a licence to a demo version of the data with

limited information content. The enterprise panel of the demo data is based on the data panel for Financial Statements statistics. The demo data can be used outside Statistics Finland to plan and test programs. These programs are run on total personal data by the staff as a charged service.

At the Research Laboratory, visiting researchers are able to link enterprise and establishment level data through encrypted unit identifiers that are similar across different data sets and over time. Information on the characteristics of personnel has also been aggregated to the establishment level, which is available to the researchers. In addition, data are available on worker inflows and outflows at the establishment level. Researchers' own data sets can also be combined with Statistics Finland's data. Appendix 2 illustrates the linking of data sets. The data are mostly in the SAS software format. In the analysis, for example, the Stata and SPSS software can also be used.

6. Future challenges

The operation of the Research Laboratory is widened continuously by increasing both the knowledge related to the data and the volume of the data. The development of documentation is an essential part of the service. The Internet gives a valuable channel for disseminating information about the facilities of the Research Laboratory to researchers.

However, increased demand for user-friendly and cost-effective access to business micro data has created the need to find alternatives to the current practices. At the moment, the service depends on time and place, which puts researchers coming from different parts of Finland in unequal positions. In addition, globalisation increases the pressure to offer the same possibilities to researchers abroad. Furthermore, the quality of research may depend on the time and financial constraints that researchers face when coming to Statistics Finland to use the data.

The future direction may be to offer the possibility to use the data on-line through a protected Internet connection. As examples, this system has been in operation in Denmark and Sweden with good experiences. One of the advantages of the system is improved data protection as use of the data can be more easily supervised and all output checked before sending them to the researcher. However, the setting up and maintenance costs of the on-line infrastructure are quite high.

The role of international and Nordic co-operation is increasingly important in the development of the Laboratory. Experiences of other countries in giving access to sensitive micro data are essential in evaluating different solutions. Furthermore, participation in international projects on micro data brings new knowledge on new targets for research and development.

Business Structures Unit
Satu Nurmi

09.10.2006

Appendix 1. Research Laboratory data sets

Enterprise level data

- The Business Register's enterprise-level statistical files (1982, 1984, 1986, 1988–2004), basic information (e.g. turnover, staff, industry) on the enterprise frame
- Group register data (1995–2003)
- Data panel of Financial Statements statistics (1986–2004), enterprises' profit and loss account and balance sheet data, key figures of financial statements
- R&D panel (1985–2004), enterprises' research and development
- Innovation data (1988, 1991, 1996, 1998, 2000, 2002, 2004), enterprises' innovation activity
- ICT panel (1998–2001), use of ICT and the Internet in enterprises
- Patent data (1985–2003), enterprises' patents
- Bankruptcy data (1986–2003), enterprises' bankruptcies
- Business Aid Database (2000–2005), business subsidies
- Other enterprise inquiries (according to the demand of the Research Laboratory)

Establishment level data

- The Business Register's establishment-level statistical files (1976, 1978, 1980, 1982, 1984, 1986, 1988-2004), basic information (e.g. turnover, staff, industry) on establishments
- Establishment panel of manufacturing statistics (1974–2004), production information of manufacturing
- Establishment-based worker and job flow data (1988–1997), establishment-level data on worker inflows and outflows according source and target
- Establishment-based data on the characteristics and pay of staff (1988–2001), e.g. pay, work experience, education and age of establishment's staff
- Establishment-based panel data on the average characteristics of staff (1988–2001)
- Establishment-based data on the distribution of characteristics and pay of staff (1988–2000)
- Inquiry on fixed assets and technology in manufacturing (1990, 2002), establishments' fixed assets, replacement value and life and use of IT in production
- Commodity statistics (1986–2004), value and volume data by establishment for products and raw materials

Individual level data

- Finnish Longitudinal Employer-Employee Data (FLEED, 1988–2003), background information on employees, which can be combined with enterprise and establishment level data

Business Structures Unit
Satu Nurmi

09.10.2006

Appendix 2. An example of linking data sets