WP. 13
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
(Tarragona, Spain, 26-28 October 2011)

Topic (ii): Software research and development

Remote access in Statistics Finland

Prepared by Janika Tarkoma, Statistics Finland

Remote access system at Statistics Finland

Janika Tarkoma[*]

[*] Statistics Finland, 00022 Statistics Finland, Finland, janika.tarkoma@stat.fi

Abstract: Remote access has been under active discussion at Statistics Finland for years. A pilot project for remote access was established in 2008 and this project continued until the end of 2009. During this time, a pilot remote access system was established with new guidelines for dissemination and scientific use of data sets in the remote access case. At present remote access is available for several research institutes and universities and the use of remote access is growing. In this paper, we describe our remote access system and the guidelines and rules for its use.

1 Introduction

Remote access has been under discussion at Statistic Finland since the 1990s. Some initial efforts for enabling Internet-based access for researchers started in 2005. However, in Finland resources allocated for enabling remote access have been limited. Finally, in 2007 we decided start the work and no longer wait for additional budget. The first aim of this project was to go through Finnish legislation concerning scientific data files and internal guidelines at Statistics Finland. Based on the research we determined the criteria for deciding on which type of data it is possible to grant access.

The second aim was to learn from other countries' systems. We studied many articles and presentations about different systems that were already established abroad and we asked the opinion of the potential users. Statistics Finland already has basic remote execution available for Finnish Longitudinal Employer-Employee Data (FLEED). This data is available on a CD-ROM version, which is a sample and protected against indirect identification. Researchers can design their statistical software programs based on this protected data, and then send them to Statistics Finland to be run on the unprotected original data. Discussions with the researchers in Finland have indicated that our users would prefer remote access because they want to have a view on the detailed microdata. Therefore, the project has focused on remote access systems.

The first phase of the project involved a state of the art survey of the existing remote access solutions developed by other national statistical institutes. These solutions have been documented in various reports, conference proceedings, and Web resources (see Hjelm, 2006; Borchenius, 2006; Hundepool and de Wolf, 2006). In addition to available documents, this phase involved also visits to institutes to learn about the remote access solutions. Most suitable systems are the ones from countries similar to Finland: small in population and registers are widely used. We observed demonstrations of Swedish and Norwegian remote access systems given by our Finnish colleagues. They were members of international research groups and had licences to use these systems. We visited Denmark and the Netherlands to learn about their remote access solutions. Sweden and Denmark have similar remote access solutions whereas the system used in the Netherlands differs from them. Our goal was to learn the technical details and the way the remote access contracts are managed. In this paper, we provide a summary of the lessons learned pertaining to the development of remote access solutions.

In this article, we describe the legal framework, its effect on microdata use, and give examples of available data sets in chapter two. In chapter three, we describe the technical details of our remote access system. In chapter four, we discuss relevant issues including possible changes in legislation in Finland.

2 Legislation and microdata access

The planning phase included a deep study of the legislation and guidelines that describe and restrict microdata release in Finland. Statistics Act is very strict even when it comes to scientific use of data sets. According to Statistics Act, "Confidential data collected a statistical authority for statistical purposes may be released for use in scientific research or statistical surveys concerning social conditions. However, personal data referred to in the Personal Data Act and the identification data of other statistical units may not be released." (Statistics Act 280/2004). In addition, our guidelines define that personal data must be protected against both direct and indirect identification even if access is granted in our research laboratory or via remote access.

Data sets are provided based on the "need to know" –principle (UNECE, 2007, p. 53). Following this principle, only those variables that are actually needed for research are included in the microdata set. This rule applies also to remote access systems. Delivering data sets via remote access is considered more secure than distributing the data on CD-ROMs. That is why; researchers may have more detailed information available via remote access system. Nevertheless, the data protection rules are the same for the different mechanisms and strong protection is needed.

Traditional way to gain access to data sets that include personal information has been specially tailored data sets (see annex 1). With tailored data sets, the researcher has defined each variable he needs and usually he has to specify a reason for the need of some sensitive variables if it is not clear from the application for the licence to use the statistical data. Then we protect a sample of the population with these variables and deliver it in CD-ROM to the user. Researchers' queuing time for microdata sets has been increasing with the increasing number of applications.

3 Remote access facility

Establishing remote access facility included the work of computer scientists, statisticians, methodologists and lawyers. Lawyers worked with contracts and helped with guidelines. Statisticians and methodologists helped deciding suitable amount of protection and computer scientists build the actual system.

3.1 Contracts

There is no legal restriction for an independent researcher to gain access to microdata for scientific use in Finland. This means that even if a researcher does not work in a university or a research institute, it is possible to get microdata sets from Statistics Finland. Researcher fills in an application for licence to use statistical data and if research plan is included, data can be provided. When we discussed remote access system, we decided that in order to gain access, users should be associated with an institute. Thus, the researchers are familiar with the rules and regulations and aware of research ethics.

There are two types of contracts used for remote access. The institute applying for the remote access has its own contract, and each research project will have a contract that the researchers (i.e., the actual users) sign. The institute has a responsible person who informs users about the rules of the remote access system. The institute is responsible that the remote access rules are followed and the institute is accountable for any misuse of the system.

Researchers agree to follow the rules. They agree not to try identification of observations and to make sure their results are in a form in which no observation is disclosed. Our personnel check a given user’s outputs before sending them to the user. Data sets accessible to the user are defined in the licences for the use of microdata sets. This is a separate from the remote access contract. The researcher must have a licence before contract can be signed.

3.2 Technical issues

We have built our remote access system on Microsoft Windows Terminal Services (TS). System is using virtual servers and VMware environment of Statistics Finland. See annex 2 for network presentation of the system. The technical description of the remote access environment is based on a Statistics Finland technical report (Verho, 2009).

3.2.1 Servers

The remote access system has nine virtual servers, which use Windows Server 2008 Standard (64 bit) as operating system and one Linux server that provides access to the storage network. The network for these servers is separated from statistics production.

TS Gateway server is focused on internet connection management. The TS Web Access service is on this server. It is used for signing into the system and logging remote access session information. The system is based on the public internet connection of Statistics Finland. The HTTPS protocol is used to secure the remote connections.

Four TS servers form the core of the system. Statistical software is located on these servers. Servers have a total of 32 GB memory so that researchers will be able to use extensive data sets. This seems to be enough for now, but when the number of the users increases, more memory will be needed.

The VMware environment distributes virtual TS servers to separate physical computers. This is a way to gain more capacity and fault tolerance for the system. TS servers are easier to maintain when we use virtual servers. It is also easier to add a new TS server to the system by cloning one of the existing servers and just adjusting its details.

Terminal Services servers include Terminal Services Log software that is using SQL-database for saving the settings. System storages log information about each user (software, processing time and memory used) in a database. TS Session Broker server is dividing burden to separate TS servers. It makes sure that when a user logs back in, he will be directed to same server where his software is still running.

TS environment requires usernames and directories to work properly and Active Directory server (AD) is taking care of these. Microdata sets and user’s folders are located on a file server. All the researchers in a research project form a user group and AD server has information about the rights each group has on microdata sets. Disk space is limited with quota definitions.

Updating and backups of the disk space for the remote access system has been integrated with the normal maintenance at Statistics Finland. System makes snapshots of the saved information from fileserver on a tape every night. It is also possible to archive research projects to tape if needed after a project is finished. A backup is created of the operating system when software is updated or at least couple of time a year.

3.2.2 Remote access system from the user point of view

A user needs a virtual desktop connection to log in the remote access system from his workspace. We use the Internet Explorer browser to connect and log in to the Terminal Service Web Access. When the user has given his username, the system will ask for a password. User receives password to his mobile phone with an SMS message. The user’s phone number has to be defined when the contracts are signed. Each password is single-use and it is sent via flash SMS, which means that it cannot be saved in the mobile phone. In our remote access system, we use strong identification where user needs id, password and some additional information that only authorized user has, like token, certificate or mobile phone SIM card.

One user can have several different projects in the system. The user chooses his research project with the TS Web Access service after he has logged into the system. When the virtual desktop is connected, the user can see a slightly modified desktop view of the Windows 2008 Server operating system. The desktop has icons for the software and available folders. The user cannot modify the view and he can only use the software on the desktop. A user has the option of several different statistical software packages: STATA, R, SPSS, SAS and normal word processing applications are available on the remote access system. It is not possible to transfer files between user's projects. In addition, it is not possible to export data from the service.

Each research project has its own work folder and everyone in the project can write to this folder. The user can add subfolders and distribute files for other researchers in the project. The user can find in his folders all microdata sets (SAS files) that his project is licensed to use. Any other data sets that the user has modified are saved to the work folder. If the researcher wants to use any of his outputs, he saves them in specified folder for output checking. Metadata files, user manuals and user rules are available in a single folder, so they are easy to find when they are needed.

When the user is finishing his work, he needs to disconnect or sign out from the session. Disconnecting will not close the software. The user can leave the software running without connection. This is very useful when the user is working with larger data sets and heavy computational tasks. The system will disconnect a session automatically if the user is not active in a predefined timeframe. The remote access software will close when the user signs out of the system.

4 Possible changes in near future

Statistics Finland has centralised facilities for researchers. This centralisation has simplified processes; however, tailor made data sets are a work intensive way to serve our customers. Standard data sets would improve processes. Data sets should include variables that are frequently requested by researchers. Our FLEED data set can be seen as a benchmark (see annex 2).

Statistics Act is to be updated in the near future. A working group has been active on this topic since October 2010 and will submit its proposal by the end of this year. If the Statistics Act is changed so that indirect identification in microdata under contract is no longer limited, this will change the principles for microdata release. Naturally, this type of data set should be provided via a secure system.

Public use files (PUF) have been under discussion. Many methodologists at Statistics Finland think that making good quality PUFs from Finnish population or companies is impossible because of small population size. Drawing a small sample of for example 1% of the population leads to sample size of some 50 000 persons which is too small. If Statistics Finland will produce PUFs, they will be intended for teaching and studying purposes.

References

Borchsenius, L. (2006). New developments in the Danish system for access to microdata. In: *Monographs of official statistics, Work session on statistical data confidentiality*, pages 13-20. Eurostat, Luxembourg.

Hjelm, C. (2006). MONA - Microdata ON_Line access at Statistics Sweden. In: *Monographs of official statistics*, *Work session on statistical data confidentiality*, pages 21-28. Eurostat, Luxembourg.

Hundepool, A. and de Wolf, P. (2006). OnSite@Home: Remote Access at Statistics Netherlands. In: *Monographs of official statistics, Work session on statistical data confidentiality*, pages 47-52. Eurostat, Luxembourg.

Statistics Act 280/2004 of Finland. Available at: <http://www.stat.fi/meta/lait/lait_statisticsact04.pdf> [Accessed 2011-9-20]

UNECE (2007). *Managing Statistical Confidentiality and Microdata Access.* UN Economic Commission for Europe, Conference for European Statistician.

Verho, J. (2009). Projektin loppuraportti (Project’s final report; in Finnish only). Reg. no. TK-04-468-08. Helsinki.

Annex 1
Examples of the data sets available for researchers

Data

Statistics Finland's enterprise data offer a diverse information basis for studying various features and development of Finnish businesses. Available are the following data and their combinations:

Research Laboratory data

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, business aid)
- average characteristics of the staff by establishment (e.g. monthly pay, education, age, gender distribution, marital status, ownership of dwelling)
- heterogeneity of the staff (e.g. distributions of pay, age and amount of education)
- staff mobility (worker flows in and out, divided by source and target)

In addition, available is also a combined employer-employee data file containing varied information about the characteristics and work history of enterprises' employees.

Enterprise level data

- The Business Register's enterprise-level statistical files (1982, 1984, 1986, 1988-2008), basic information (e.g. turnover, staff, industry) on the enterprise frame
- Group register data (1995-2008)
- Financial statement data panel (1986-2005, of which EVR 1994-2005), enterprises' profit and loss account and balance sheet data, key figures of financial statements
- R&D panel (1985-2008), enterprises' research and development
- Innovation data (1988, 1991, 1996, 1998, 2000, 2002, 2004), enterprises' innovation activity
- ICT panel (1998-2001, 2001-2005), use of ICT and the Internet in enterprises
- Patent data (1985-2004), enterprises' patents
- Business Aid Database (2000-2008), business subsidies
- Other enterprise inquiries (according the demand of the Research Laboratory)

Establishment level data

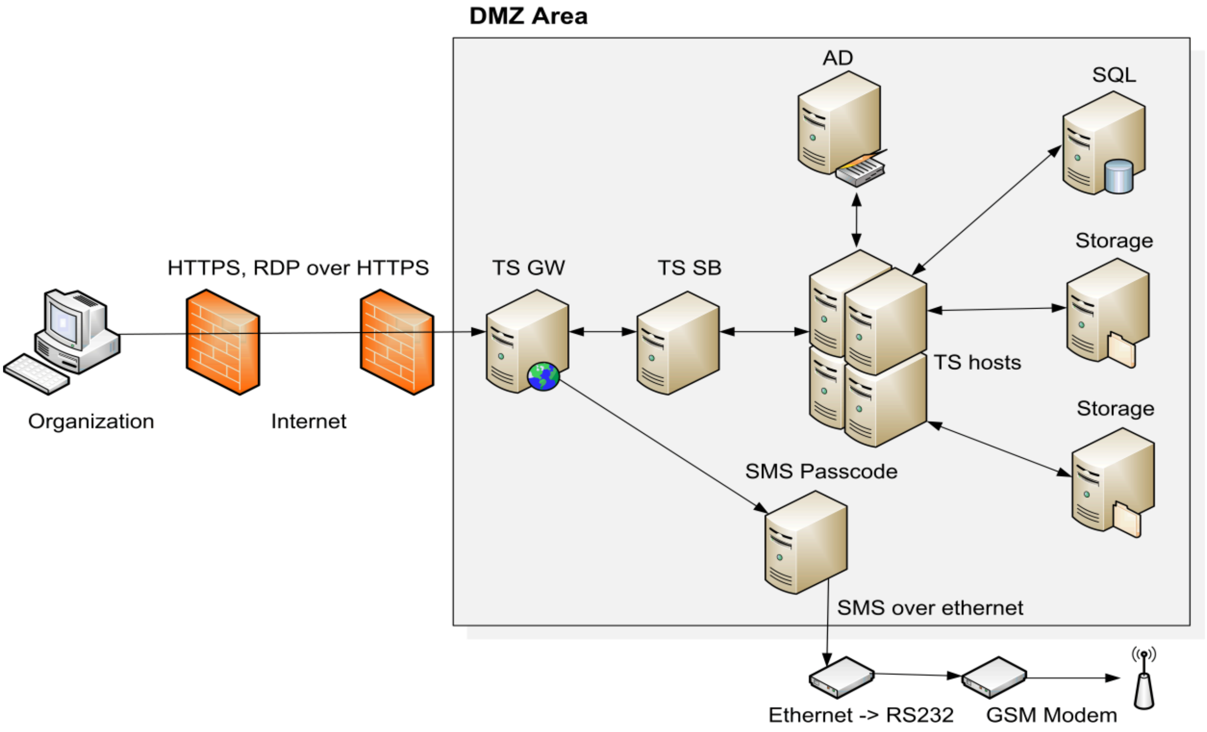
- The Business Register's establishment-level statistical files (1976, 1978, 1980, 1982, 1984, 1986, 1988-2008), basic information (e.g. turnover, staff, industry) on establishments
- Establishment panel of manufacturing statistics (1974-2007), production information of manufacturing
- Establishment-based worker and job flow data (1990-2005), establishment-level data on worker inflows and outflows according source and target
- Establishment-based or enterprise-based data on the characteristics and pay of the staff (1988-2007), e.g. the pay of the establishment's staff work experience, education, age
- Inquiry on the fixed assets and technology of manufacturing (1990, 2002), establishments' fixed assets, replacement value and life and use of IT in production
- Commodity statistics (1986-2008), value and volume data by establishment for products and raw materials

Individual level data

- Finnish Longitudinal Employer-Employee Data (FLEED, 1988-2007), background information on employees, which can be combined with the enterprise and establishment level data
- Available are also FLEED demo data with limited information content. The enterprise panel of the demo data is based on the financial statement panel. The demo data can be used outside Statistics Finland for planning and testing programs.

In addition, the Research Laboratory has certain industry-specific files, such as producer prices and OECD's STAN database, and conversion keys for industries and for unification of municipalities.
Available at: <http://tilastokeskus.fi/tup/yritysaineistot/aineistot_en.html>
[Accessed 2011-9-20]

Annex 2
Network presentation of the remote access system


DMZ Area
AD
SQL
Storage
HTTPS, RDP over HTTPS
TS GW
TS SB
TS hosts
Storage
Organization
Internet
SMS Passcode
SMS over ethernet
Ethernet -> RS232
GSM Modem