Riitta Harala
Anna-Leena Reinikainen

Seminar on the principle of statistical confidentiality, 13 January 2005, Lisbon

Data confidentiality in the processing of personal data for statistical purposes in Finland

1. General

The processing of personal data for the purposes of official statistics is in Finland subject to the provisions of the Finnish Constitution (2000), the Finnish Personal Data Act (1999), which is based on the Data Protection Directive of the European Union, and the Finnish Statistics Act (2004).

According to the Constitution, the right to privacy is a basic right and detailed provisions on the protection of personal data are laid down by law. The Constitution has been interpreted as meaning that the norms of legislation on the protection of personal data may not fall below the standards set in it. The Constitution has also been seen as requiring that any legal provisions concerning the processing of personal data must be highly detailed.

The Personal Data Act applies alongside the Statistics Act to the processing of personal data for statistical purposes. With regard to the compilation of statistics and scientific research the Personal Data Act contains several exceptions to its general provisions. These are based on the Data Protection Directive of the European Union. The Act allows utilisation of data collected for other purposes, such as administrative ones, in the compilation of statistics. It also allows the collection of sensitive data for statistical purposes. It does not oblige statistical authorities to arrange the right for the persons entered in personal registers to check the data concerning themselves, because the data in the registers are not used for making decisions that affect their benefits or rights. It is sufficient if a general description of the contents of the data in personal registers is made commonly available. The descriptions of the personal registers of Statistics Finland are commonly accessible via the Internet.

The Statistics Act lays down provisions for the procedures and principles concerning the collection of data and the designing and production of statistics that state authorities compiling them must abide by. The provisions of the Act cover the stages of the statistics compilation process from the collection of data to their release or publishing. The Act puts emphasis on the rationalisation of data collection, creation and maintenance of good relationships with data suppliers, provision of information to data suppliers, and observance of the principles of data protection and statistical ethics.

Statistics Finland’s production of statistics is characterised by broad exploitation of data in administrative registers. This exploitation of register data was started as far back as the 1970 census of population and dwellings, and has since been continued and widened from one year to the next, first in the area of population statistics and then also in business statistics. In 1990, Finland was the second country in the world after Denmark that was capable of drawing a population and dwelling census entirely from administrative registers, and has since produced practically all population census data annually with the help of its register-based system.

Data in administrative registers can only be exploited efficiently if different registers use uniform identification coding systems. Uniform personal identity codes were introduced in Finland in 1963 and these are used in all administrative registers of individual persons. Besides the personal identity codes, there are also other, almost equally consistent coding systems in Finland for e.g. enterprises, and buildings and dwellings, and these facilitate the exploitation of registers in the production of statistics.

At the moment Statistics Finland is revising the production model of its statistical systems within a wide-ranging project, and data protection solutions will receive special attention in this context. The objectives of the revision are standardisation of the production model and digitalisation of the entire process from the collection to the dissemination of data, as well as improvement of data storage and exploitation by adoption of the data warehouse approach. All mainframe-based large systems for population statistics, including those within the body of the register-based census system, will be reviewed in connection with this revision, and their production will be moved to the open environment.

2. Treatment of personal data in data collection

According to the Statistics Act, when data are collected for statistical purposes the primarily exploited sources should be the data accumulated in administering the tasks of general government and those produced as a consequence of the normal activities of employers, self-employed persons, corporations and foundations. This means that when data collections are being planned, possibilities of drawing the required data from these sources must always be explored. A separate collection of data for statistical purposes may only be started if the information cannot be obtained in this way. If a new collection is made, the authority producing statistics must see to it that the respondents are only requested to provide those data that are necessary for the production of statistics. Furthermore, the data must be collected in a manner that is economical and causes the respondents the least amount of inconvenience.

The Statistics Act contains provisions on the obligation to provide data for Statistics Finland. In respect of personal data, the obligation to provide data is covered by a separate section of the Act and is prescribed in more detail than the collection of other data. The Act lists the topics which the obligation to provide personal data might concern. Furthermore, it specifies when sensitive data may be collected. The obligation to provide personal data applies to state authorities, entrepreneurs and other employers, certain central pension institutions, and organisations providing education. Because personal data are largely obtained from the files held by these parties, no legal obligation to provide data is imposed on individual citizens from whom data are always collected on voluntary basis.

The point of departure in the Statistics Act is that data are collected and stored without identification data whenever permitted by the statistics to be produced. Identification data, such as personal identity codes, names or addresses may only be collected where it is necessary for the linking of datafiles or otherwise essential for the production of reliable and comparable statistics.

In practice, efficient utilisation of existing administrative data as required by the Statistics Act always necessitates use of the personal identity code when unit level data are being collected on individuals for register-based statistics, or as auxiliary data for statistics compiled from data collected by interviews or questionnaires. The personal identity code acts as the link when data from different registers are combined with other data.

Collection of data from administrative registers

Approximately 95 per cent of the data of Statistics Finland today come from administrative sources while the remaining 5 per cent are collected direct from individual persons, enterprises, corporate bodies and educational institutes. Personal data are collected both from administrative sources as well as direct from individuals themselves with interviews or questionnaires, or indirect from educational institutes on their students and teachers and from employers on their employees.

While considering the exploitation of administrative register data and data protection it is good to bear in mind that the flow of information runs in one direction only, i.e. **from the administrative authorities to the statistical institute – never the other way round**. The administrative authorities would certainly benefit from much of the information collected and linked for statistical purposes, but the statistical institute must never release the outcomes back into administrative systems.

Although legislative obstacles to the statistical use of administrative sources are sometimes justified by reasons of data protection, register-based statistics also offer various **advantages in terms of data protection**. In register-based statistics production, the number of people who have access to the statistical data is much smaller than in traditional data collection. For example, the plain-language questionnaire forms collected in the 1980 population census were spread out for several months in 350 regional offices. Some 2,200 short-term employees were involved in processing these forms, and any information on their neighbours might well have aroused their interest. It is clear that in this situation it is much harder to maintain high standards of data protection than it is with register data, which come to Statistics Finland in machine-readable format and are shown to no one else than the computer. It is much easier to keep account of a magnetic tape, a cassette or a hard disk datafile than it is of miles of archives spread out across the country.

Another definite data protection advantage of register-based census collection is that it means there is no further need for subcontractors. Subcontractors were needed and used in virtually all questionnaire-based population censuses in Finland: after all these involved mailing and recording data from millions of forms within a short space of time.

Administrative data come to Statistics Finland from a variety of authorities in diverse forms, e.g. by direct line transfer, on magnetic tapes or diskettes or as email attachment files. As a rule, the data come equipped with personal identity codes. Direct line transfer takes place in a protected connection via a separate line. Magnetic tapes and diskettes are brought to statistics Finland as personal deliveries or by post, when the data are covered by legislation on the secrecy of correspondence.

Direct collection of data

Data are collected **direct from individuals** in various sample surveys, such as the labour force survey, household budget survey, leisure survey, time use survey, and so on. These data are collected with either interviews or questionnaires. Statistics Finland has its own interviewer organisation of around 150 people who work in different locations across the country. The interviewers have laptop PCs into which they collect data. In these data the identification data of individual persons are represented by so-called target numbers, which link the interviewer's data with personal identity codes on separate target questionnaires. The interviewer sends the interview data from his or her own PC as direct line transfer to Statistics Finland via a protected connection. The target questionnaires are returned to Statistics Finland by post. At Statistics Finland, personal identity codes are restored to the data for the linking of registers, after which they are removed again for the duration of data

processing. However, the key between a target number and a personal identity code is retained for possible future needs.

Correspondingly, questionnaire forms are not pre-filled with personal identity codes but questionnaire numbers act as links to them for the combining of files that is done at Statistics Finland.

Personal data are collected **indirect** from educational institutes, municipal authorities and private enterprises. Electronic data collection is increasingly being used for indirect collections. Diskettes or email are still used in some collections but are replaced more and more by collections via the Internet. Two models are used for collecting data via the Internet, one developed by Statistics Finland itself, and one in which data are collected via an external operator. Up to now, the latter has proven better suited for the collecting of personal data and it is used in most collections targeted at educational institutes. In it, data suppliers fill in a www questionnaire in a service provided for them by the operator and then send it back via a protected connection to the operator who then gathers the questionnaires together and transmits them on to Statistics Finland. Access to the connection requires user names and passwords. The operator is bound by the same legal provisions concerning data protection as Statistics Finland.

Personal data are also still collected indirect with questionnaires, especially from enterprises. These data concern the establishments and occupations of employees of multiple establishment enterprises. Enterprises record the data either on a questionnaire, or on a diskette or some other recording device and send these by post, whereby the data are covered by legislation on the secrecy of correspondence.

The aim within the project to revise the production model of Statistics Finland's statistical systems is to construct an electronic alternative for all data collections by the end of 2006. The model will either be one developed by Statistics Finland for data collections via the Internet or one based on the use of an external operator as described above. A new, Citrix-based data transfer system is also being planned for the transmission of data collected with interviews to Statistics Finland. In the new system data would no longer be stored on laptop PCs but the interviewer would contact a computer at Statistics Finland and store the data direct on it. This would improve data protection considerably.

3. Handling of personal data during processing

The Statistics Act requires that when data collected for statistical purposes are being combined, stored, destroyed or otherwise processed, it must be seen to that no person's privacy, or business or professional secret is endangered. In addition, it must be seen to that the data are duly protected during all stages of statistics production. The Act also requires data processing to take place in accordance with statistical ethics, good statistical practice and international recommendations and procedures generally applied in the field of statistics.

An essential part of taking care of the protection and security of data in confidential statistical files is making sure that confidential personal data are only processed by those whose tasks demand it. In practice, access to the basic data of different statistics is administered with user rights. User rights are granted on the principle of least authority, in other words setting out from the rule that only persons participating in the compilation of certain statistics may have access to the data in the related statistical system. User rights to the data of each statistical unit of Statistics Finland are granted by the director of the unit concerned. There are six statistical units: Population Statistics, Social Statistics, Prices and Wages, Business Structures, Business Trends and Economic Statistics. In practice, the directors of the statistical units have broadened the user rights to some extent within their own units, especially in

respect of statistical systems in which datafiles are frequently combined for charged commissions or research purposes. A written licence application procedure is used for granting user rights to the data of a statistical unit other than own unit. As a rule, such licences are granted for a fixed term.

Taking, processing or storing confidential unit level data outside the agency's premises is not allowed. Likewise, distance use of basic data in connection with e.g. teleworking is disallowed.

As a component of the revision of Statistics Finland's production model a project will commence this year in which a new system will be built for the administration of user rights that will have an improved automatic ability to recognise transfers of employees between tasks. Similarly, a more specified form of procedure will probably also be adopted in the granting of user rights within the statistical units.

When the major systems for producing population statistics, such as demographic statistics, family statistics and employment statistics, are being redesigned in connection with the production model revision provisions are also being made to enable the processing of the majority of personal data with changed personal identity codes. A key for linking them to the original personal identity codes will be retained.

4. Personal data and data dissemination

Irrespective of the source, data obtained for statistical purposes are confidential with the exception of the specific cases itemised in the Statistics Act, such as certain data in the Register of Enterprises and Establishments, and those describing the activities of central and local government authorities and the production of public services. Breaches of statistical confidentiality are punishable under the Penal Code, so that persons serving a statistical authority may be sentenced to a fine or a maximum of two years' imprisonment, and other persons to a fine or a maximum of one year's imprisonment. Up to now, no breaches of statistical confidentiality have occurred.

Confidential data may only be released to a third party on terms laid down in the Statistics Act or in another act concerning specifically the National Statistical Service. Thus, the obligation on confidentiality prescribed in the Statistics Act cannot be repealed by any other act. Furthermore, the Statistics Act expressly forbids the use of data in an investigation, surveillance, legal proceedings, administrative decision-making or similar handling of a matter concerning an individual, enterprise, corporation of foundation.

According to the Statistics Act, data may only be released for use in scientific research or for statistical surveys on social conditions in such a form that individual persons cannot be identified direct or indirect from them. Exceptions to this are data on age, gender, education and occupation, which Statistics Finland may release inclusive of identification data for use in scientific research or statistical surveys. These latter mentioned data are mainly released with identification data for research in the health field.

The Statistics Act requires that compiled statistics be as reliable as possible and give a truthful picture of social conditions and their development. Statistics must be compiled so that those whom they concern are not directly or indirectly identifiable from them unless the data concerning identification are public by virtue of the Statistics Act. Statistics must also be published as soon as possible after their completion.

Protection of data in table format

In 2002, Statistics Finland drew up guidelines for the protection of **personal data in table format**. In the guidelines personal data are classified into three categories dependent on the strength of data protection measures they require as follows:

1) As a rule, statistics containing the least sensitive basic variables, such as demographic data, can be released in table format, and as figures or map presentations without special data protection measures.

2) It is recommended that statistics containing somewhat sensitive variables, such as income data, should be protected by applying the threshold value of at least five to the persons or household-dwelling units being classified. However, the classifications must be selected so that no table cell or as few of them as possible contain only one unit.

3) Statistics describing sensitive variables, such as race or political or religious conviction, must always be compiled from data on large populations and regions so that the threshold value in any one table cell will be at least 10. User licences should also be considered before releasing statistics containing such variables.

Besides the sensitivity of the variables, the protection needs of the data in a given table are also determined by the number of observations per table cell, size of the studied population, number of variables, and size and ease of locating the geographic area to which the data relate. Further matters that should be considered when assessing data protection needs are whether the values in the table are absolute or relative, the data old or new, or concern one year or represent the sum or average for a number of years.

At the moment, the most problematic aspect in administering the protection of table format data is that there is no good data protection tool that could be linked up to the tabulation process, and individual tables still demand certain amounts of manual work. The protection of data on small areas is another subject that needs consideration. It is possible to produce information by quite precise areas (e.g. 250mx250m map squares) thanks to the availability extensive statistical datafiles based on map co-ordinates, but in a sparsely populated country like Finland such data must often be disguised for data protection reasons.

Release of microdata for research purposes

By virtue of the Statistics Act, Statistics Finland releases annually approximately 200 sets of **microdata** to outside researchers. A vast majority of these consist of personal data. In addition, a few dozen researchers make use of Statistics Finland's research laboratory to study materials that for reasons of data protection may not be released outside the agency. Most of these are data on enterprises, in some cases combined datafiles on enterprises and employees.

Released microdata are always edited into an unidentifiable form. In addition to the removal of all identification data they are also otherwise edited to prevent indirect identification. As a rule, Statistics Finland does not release total datasets for research purposes. In this context total datasets refer to populations containing all persons possessing a certain characteristic, e.g. all persons having attained a certain educational qualification, living in a certain municipality, and so on. The level of detail by which variables are classified may also need to be reduced or certain combinations of variables removed in order to prevent identification. Data protection procedures are determined for each dataset case by case. In order to unify practices, the procedures for protecting microdata will be gathered into a set of guidelines during this year.

User licences are granted to research data by way of an application procedure. A user licence application must specify the purpose for which the data will be used, the persons who will participate in their processing, the data to which user licence is applied for and the duration of their usage.

Decisions on the release of statistical datasets for research purposes are made by the respective directors of the statistical units. Difficult cases are submitted to Statistics Finland's statistical ethics committee, which consists of representatives of various statistical units. Decisions on the release of research materials to foreign countries are made by the Director General after consultation with the ethics committee.

In order to rationalise the use of research material and improve data protection Statistics Finland will investigate the possibility of building an online facility, based on the previously mentioned Citrix system, in which the material would not move outside the agency's premises but could be teleused from a distance. This kind of a solution is already in use in Denmark and has just been launched in Sweden.

The demand for data for research purposes has been growing continuously in recent years as awareness of this possibility has spread thanks to Statistics Finland's own active measures and researchers' own experiences. As researchers have become increasingly aware of what kind of data are available, the applications for user licences have also become more complex and expansive. At the same time, taking care of data protection has grown more and more demanding. A statistical authority must constantly try and find a balance between the data needs and demands of researchers, optimum exploitation of statistics in society, and making sure that suppliers' data remain protected. Finding this balance is of primary importance.

5. Summary

Work to maintain the highest possible level of data security and public confidence is an ongoing effort for every statistical authority. At Statistics Finland, these efforts are most clearly evident in the work of the statistical ethics committee and the data protection working group. However, the work essentially entails taking care of the flow of the normal, daily statistics production process. In the past year, Statistics Finland has focused strongly on its data security policy and on determining the consequent measures, as well as on reviewing its guidelines on data security. This work will continue through the current year. Personnel training in data security, aimed at bringing data security matters to the fore, has also been planned for this year. Data security is often also a matter of attitude; it is not enough to have technological solutions and guidelines in order; people must also use and abide by them.

In terms of technology, Statistics Finland also has great expectations as regards data protection and security from the work being done to revise the production model of its statistical systems. The revised model based on the uniform data warehouse approach makes Statistics Finland's internal administration of user rights even more important than before. On the other hand, the growing movement of data via information networks, both in their collection and dissemination, highlights the importance of secure transmission solutions and user right administration also from the perspectives of the suppliers and users of data.

References

Harala, R, and Reinikainen, A-L. (1994), 'Statistical confidentiality and the use of statistical data for research purposes – Finnish aspects', Second International Seminar on Statistical Confidentiality, Luxembourg 1994.

Harala, R, and Reinikainen, A-L. (1996), 'Confidentiality in the use of administrative data sources', Statistical Journal of the United Nations Economic Commission for Europe, 4/1996.
Statistics Finland (2004), Use of Registers and Administrative Data Sources for Statistical Purposes, Best Practices of Statistics Finland, Handbooks 45, Helsinki 2004