United Nations

ECE/CES/2013/37

Economic and Social Council

Distr.: General
27 May 2013

English only

Economic Commission for Europe

Conference of European Statisticians

Sixty-first plenary session
Geneva, 10-12 June 2013
Item 4(b) of the provisional agenda
How should national statistical offices respond - moving from risk avoidance to risk management

Micro-data services in the Netherlands

Note by Statistics Netherlands*

Summary

This paper describes the process of providing access to micro-data for research purposes, including the kind of data made available, the security measures put in place, the services provided, the related fees and contacts between researchers and Statistics Netherlands. The paper also considers how to develop the process in the future. Finally, the paper provides a vision of possible international developments and a summary of the main conclusions.

* This document was submitted late due to delayed inputs from other sources.

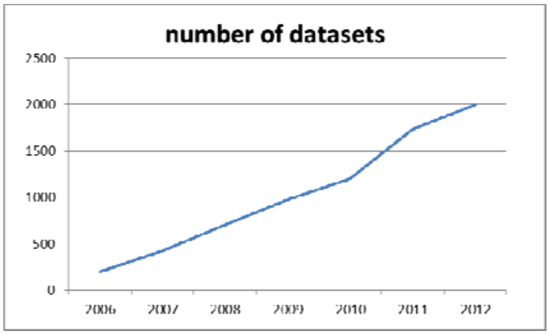
I. Introduction

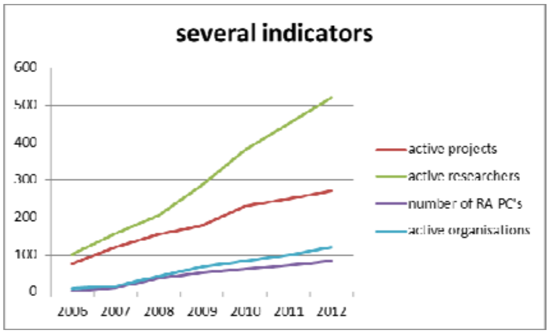
1. More than ten years ago, Statistics Netherlands started providing access to secure-use files for the purpose of scientific statistical research. This service is now an indispensable part of the services provided by Statistics Netherlands. Organisations, such as the Netherlands Bureau for Economic Policy Analysis and various research agencies, would not be able to do their work efficiently without this service. The users of micro-data also include many universities.

2. Although these services are part of Statistics Netherlands' work programme they are not obligatory nor are they incorporated in the regular statistical production. Secure-use files collected in the course of the statistical production are made available 'as is'. However, the meta-data are considerably more extensive than is necessary for internal use.

3. In the last ten years, the service has been expanded step-by step (see Figure 1), and more than 2000 well-documented datasets are now available for research. As the number of available datasets increases, usage of our services also increases. More than 100 organisations within and outside the Netherlands work with our secure-use files, with a total of more than 500 registered researchers. They come from universities, policy analysis institutes, academic research institutions, ministries and municipal statistical research centres and statistical research agencies. Researchers can work at one of the eight on-site work stations at Statistics Netherlands, or use remote access (RA) facilities. There are now 85 such RA locations, including several outside the Netherlands (in Germany, Italy, Denmark and the United States). Every month 10 new projects are added and there are over 280 active projects at this moment.

Figure 1
Expanding the service of providing micro-data sets for researchers





II. Data

4. There is a considerable lack of clarity concerning terminology in the international discussion surrounding the provision of micro-data. Different types of organisations view data from different perspectives. The resulting confusion hampers discussion. The list of terms compiled recently by the OECD[1] only exacerbates this problem. In this paper, we use the following terms[2]:

(a) Raw data files contain all replies by each respondent obtained immediately after data entry;

(b) Confidential data for scientific purposes are data which only allow for indirect identification of the statistical units, taking the form of either secure-use files or scientific-use files;

(c) Secure-use files contain confidential data for scientific purposes to which no further methods of statistical disclosure control have been applied. These are data at the level of the statistical unit;

(d) Scientific-use files contain confidential data for scientific purposes to which methods of statistical disclosure control have been applied to reduce to an appropriate level and in accordance with current best practice the risk of identification of the statistical unit.

[1] Organisation for Economic Co-operation and Development (OECD) Expert Group for International Collaboration on Micro-data Access (2012), STD/CSTAT/MICRO(2012)5

[2] Based on Regulation (EC) No. 223/2009 of the European Parliament and of the Council on European Statistics as regards access to confidential data for scientific purposes

5. One important additional issue concerns the possibility to link the secure-use files to each other. In the datasets provided by Statistics Netherlands, although statistical units cannot be directly identified, unit X in dataset A is the same as unit X in dataset B. This increases the risk of indirect identification, as the more details that are available about statistical unit X, the easier X is to identify. Linkable data are much more valuable than separate non-linkable data, but they also involve a much greater risk of indirect identification.

6. Statistics Netherlands does – to a limited extent - compile scientific-use files and public use files (PUFs). These are transferred to the national data archive for scientific research called Data Archiving and Networked Services (DANS)[3]. In addition to data from Statistics Netherlands, DANS offers numerous other datasets for scientific research. Making these datasets involves a lot more statistical disclosure measurements than secure-use files. The datasets are not linkable.

7. Dutch official statistics are published in publications and via Statistics Netherlands' website (StatLine)[4]. StatLine is a dynamic system which enables users to generate tables, graphs and maps based on statistics. The statistics themselves can be downloaded and are in fact PUFs.

III. Security

8. The Act on Statistics Netherlands[5] states that the following organisations can be granted access to secure-use files:

(a) Dutch universities (under the Law on Higher Education and Scientific Research);

(b) A statutory organisation or Dutch institute for scientific research;

the Netherlands Bureau for Economic Policy Analysis (CPB); the Netherlands Institute for Social Research (SCP); the National Institute for Public Health and the Environment (RIVM); the Netherlands Environmental Assessment Agency (PBL);

(c) Eurostat: national or community statistical authorities of the member states of the European Union.

9. Institutions or organisations not included in one of the above groups may be authorised by the Central Commission for Statistics (CCS). However, they must comply with the following conditions: the institution must be independent and have an independent legal personality, the research performed by the institution must be independent and scientific, and the institution must provide public access to its published results. These organisations, such as statistical research bureaus commissioned by a third party (usually the government) do not receive general authorisation. They are required to apply for a separate authorisation for each project.

10. Once authorisation has been granted, the research proposal is assessed, to verify which datasets are required for the project. Access is limited to the datasets that are required for the research at hand.

11. All researchers working on the project are required to have an employer's statement proving that they are actually employed by the organisation that has been granted

[3] www.dans.knaw.nl

[4] http://statline.cbs.nl/statweb/

[5] http://www.cbs.nl/NR/rdonlyres/BBD8113D-7EE5-4BE4-8879-685253B31882/0/statlawen.pdf

authorisation. They are also all required to sign a confidentiality declaration and to have their fingerprints scanned.

12. Researchers access the data via a secure internet connection (Citrix). In the case of remote access, the personal computer (PC) used must be placed in a secure location. Among other conditions it must be located in a separate room that can be locked when the PC is not in use. A fingerprint reader is connected to each remote access-PC. Researchers have to identify themselves with a fingerprint to log in.

13. Researchers cannot download or e-mail data into or from the project environment. All results, including results of calculations, remain on the disks at Statistics Netherlands and do not leave our organisation. Results or tables that a researcher wishes to use in his own working environment will be scanned by Statistics Netherlands with regard to disclosure risks. Over 500 of such output checks are carried out every year.

14. All research projects are treated equally. The same security measures apply to everyone. However, not all data are treated in the same way. Some datasets are more sensitive than others, for instance data on causes of death, and on crime suspects and convicts. Extra conditions of use apply for these datasets.

15. Especially once users are used to the provisions and start taking them for granted, constant attention for security is an important component.

IV. Standard services

16. New clients first contact the front office which consists of five account managers. If necessary, the client receives help and instructions about the application and authorisation process. They are subsequently assisted with the process of confidentiality declarations, fingerprinting and the employer’s statement.

17. Clients who want to install a remote access-PC have to acquire their own PC and create a secure location for it. A member of Statistics Netherlands staff will then install the fingerprint reader and software, and check that the location complies with security requirements.

18. Each project starts with an intake interview during which the required datasets for the research proposal and the necessary statistical programs are discussed. This step can be omitted or conducted by e-mail for experienced clients. The datasets available from Statistics Netherlands are listed in a catalogue that is available online[6]. The datasets are well documented and include metadata. In our experience the metadata required for the statistical processes of Statistics Netherlands are not sufficient for external researchers. If a researcher needs datasets that are not yet listed in the catalogue, datasets may be compiled especially for the study.

19. A project environment is subsequently designed in which datasets are ready for access. The standard environments we provide comprise 40, 60, 80 or 200 megabytes (MB) disk space, and Excel, the Statistical Package for the Social Sciences (SPSS) and the programming language R for processing. If necessary, Statistics Netherlands can provide more MBs or other software, such as the Statistical Analysis System (SAS) or the Data Analysis and Statistical Software (STATA).

[6] http://www.cbs.nl/en-GB/menu/informatie/beleid/catalogi/catalogus-micro-databestanden-thema/default.htm

20. During the project we provide a helpdesk for technical problems and questions about the datasets. In the course of the research users can add researchers, datasets from the catalogue and can import new software, datasets or activate other statistical programs.

21. Output is supplied to Statistics Netherlands and is checked for disclosure risks within two working days. The quality of the output is not checked. That is the responsibility of the researchers themselves.

22. When the project has been completed, the datasets, developed programs and results are stored for five years. Clients can extend this period if required.

23. We have recently introduced a new service: Direct Data Access. This is intended for users (mainly employees of government ministries) who have to do research within a very short period, for example to answer questions in parliament. For these purposes a one-off project is set up with specified datasets and researchers. These are normally not activated but can be activated within 24 hours if necessary, so that the researcher can start work on the datasets immediately. Output checks for these projects are given a priority.

24. Information on the services we provide is available online in a services catalogue[7]. This includes throughput times and fees as the services are not provided free of charge.

V. Costs

25. Although the data are free, services related to providing access to them are not. The services providing micro-data access are not necessary part of the regular work programme of Statistics Netherlands, and are therefore not covered by the regular budget.

26. To make the secure-use datasets available in total 24 blade servers have been installed, with a storage capacity of more than 4 terabytes (TB). Special network provisions are also necessary including the Citrix licences. Licences for the programs used for processing also have to be paid for, e.g. SPSS, MSOffice, STATA, SAS, geographic information systems (GIS), WinZip, Acrobat Reader and others.

27. Account managers, 4 full time equivalents (FTE), are active in the front office. Dedicated back office staff work on documentation, updating the data catalogue and compiling specialized datasets (4 FTE). Furthermore, information technology (IT) management and administration (5 FTE) and coordination (1FTE) work are needed.

28. In total, at the present volume, the provision of micro-data services costs just over 2 million euros per year. Statistics Netherlands subsidises this with 700 thousand euros. The remainder (1.3 million euros) has to be recovered from user fees. Statistics Netherlands has, therefore, calculated the total cost of these services, and will charge part of these costs from users.

VI. Contact with researchers

29. As Statistics Netherlands aims to fit its secure-use data services to the wishes and demands of researchers as much as possible, it is important to keep in touch with the researchers. The front-office is frequently in touch with researchers at the start of and during the projects.

[7] See: http://www.cbs.nl/en-GB/menu/informatie/beleid/zelf-onderzoeken/default.htm

30. In addition, researchers are represented in a user council which meets three times a year. It is chaired by one of the users and the secretariat is operated by Statistics Netherlands. This council discusses policy changes and changes in tariffs. Although the user council has a lot of influence, Statistics Netherlands remains responsible for the decisions made. Reports of the council are made available to all researchers.

31. Following each project, and in the case of long-term projects, during the course of work, a user satisfaction survey is conducted, in which researchers are asked how satisfied they are with a number of aspects of the process. This provides important input for improvements to the services. On a scale of 1 to 10 the overall score is currently 7.4.

32. Once or twice a year Statistics Netherlands organises a meeting for all researchers. During these meetings, we provide information about the services and about relevant changes. Researchers are also informed of future plans and they can exchange experience and knowledge with each other.

33. Part of Statistics Netherlands' website is especially reserved for secure-use data services, including the related catalogues. In the future we plan to add a section on frequently asked questions (FAQ) and a wiki environment. This facilitates creating a community in which researchers can provide input and exchange experience and knowledge. Obviously this will also help Statistics Netherlands see where the researchers experience problems and thus improve the service.

34. Lastly, Statistics Netherlands publishes a newsletter two to three times a year, which is sent to all researchers. This contains information on all the major new developments.

VII. Future developments in the Netherlands

35. Although Statistics Netherlands has been providing secure-use data access for over ten years now it continues to develop new services. In the future it shall add more and more datasets generated by the statistical production processes at Statistics Netherlands. Statistics Netherlands increasingly uses register data instead of sample surveys. The number of datasets that give a complete picture of Dutch society will, therefore, increase. This provides more possibilities for research, especially because these datasets are linkable. Security risks, however, are also much bigger.

36. The services Statistics Netherlands provides continue to evolve; Statistics Netherlands is always looking for new services and new options for the researchers. For example, benchmarks of similar services in other countries are carried out, hoping to find possibilities for implementing some of them. Unfortunately, the results of this benchmarking exercise are not yet available.

VIII. International developments

37. International exchange of secure-use files has been a topic for a very active discussion for a number of years. The Organisation for Economic Co-operation and Development (OECD), United Nations, Eurostat and the European Union (EU) all have projects in progress aimed at promoting this exchange. At European scale these projects are aimed at a centrally managed system in which secure-use files can be used simultaneously for scientific research.

38. Because participants in these projects, experts groups and forums include not only national statistical offices (NSOs), but also data archives, discrepancy in definitions is often an issue. Most recently, this has resulted in Regulation No 223/2009 of the European Parliament and of the Council of European Statistics as regards to access to confidential

data for scientific purposes. This not only comprises clear definitions of types of secure data (as stated above), it also lists the research entities eligible for access to secure-use files at EU level, the conditions which a project proposal must fulfil, the role and position of the NSOs, and the facilities to be developed to enable all this.

IX. Conclusion

39. Making secure-use files available or providing access to them has added value for society. This is borne out by the Dutch situation. Policy analysis institutes have indicated that their work would be severely hampered without access to micro-data. The volume of research and the range of subjects of studies using these data are substantial. Much of the added value for society was not anticipated: more parties than foreseen see opportunities for research on more topics than anticipated. It, therefore, does not come as a surprise that there is international interest in making secure-use files available to other countries. The experiences in the Netherlands can help to address some of the following points.

40. The security of these highly sensitive datasets needs to be an important principle. Granting access to secure-use datasets requires a combination of trust and privacy. Security measures have to be drawn up. The risks for NSOs are great. Although secure-use data services are a by-product of Statistics Netherlands' statistical production process, abuse of the trust and privacy principles would severely damage its public image. The secure-use datasets, therefore, never leave the computer systems of Statistics Netherlands. How can these principles of trust and privacy be upheld in an international system? Will datasets with individual data on the income of sport stars, pop starts or captains of industry be made available? Can datasets containing information about companies be secured against industrial espionage? Who will be responsible for output checks?

41. It is quite complicated to get the system for secure micro-data access up and running. It requires a lot of work. Actually making the data available is only a small part. An organisational structure, processes that makes sure everything goes smoothly and an IT environment that makes it all possible is only a skeleton of what is required. Who will organise this at European level? Which part of the organisation should be central, which part should operate at national level?

42. It costs money to provide access to secure-use files. Although it is tempting to state that data are free, the service comprises more than data alone. To date, the cost aspect has received too little attention in the international discussion, and this issue will have to be addressed in more detail in the development stage as the amounts involved are considerable. Who will pay for provision of micro-data at international level? Will researchers have to contribute to these costs? How will the revenues be distributed among the NSOs to cover their costs?

43. There is a certain amount of confusion in the area of definitions concerning just how sensitive datasets are. An internationally sanctioned list of definitions should be drafted as soon as possible. NSOs should take the lead in this respect as they have most to lose from ambiguous definitions. The definitions in the EU Regulation are a good start, but they must be worked out in further detail into practical definitions which specify the level and method of statistical security. There is, for instance, insufficient clarity on the issue of linkable datasets. Is it the intention to use secure-use datasets at international level? Are linkable datasets foreseen?

44. Although national legislation differs from country to country, NSOs appear to grant access to secure-use files only under very strict conditions, and in most cases the data itself - or authority over the data - may not be transferred. Collecting secure-use files at EU level clearly contravenes this principle. The autonomy granted under Regulation 223/2009 (i.e.

The approval of the national statistical authority which transmitted the confidential data concerned shall be sought for each research proposal before the access is granted[8]) assumes that confidential datasets have already been supplied to a central system. Technological provisions such as distributed computing make it possible to do research with secure-use files without these having to be centrally stored. Are these developments adequately addressed in the international discussion?

45. If the Dutch experience can be taken as a measure for the international situation, it is obvious that these problems can be overcome, however not easily. The added value of overcoming them for society is so large, that it will certainly be worth it.

[8] Article 6 of the regulation mentioned above