Distr.
GENERAL

ECE/CES/SEM.54/13
12 May 2006

ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES (EUROSTAT)**

**ORGANIZATION FOR ECONOMIC COOPERATION
AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE**

Joint UNECE/Eurostat/OECD Seminar on the Management of Statistical Information Systems (MSIS)
Sofia, Bulgaria, 21-23 June 2006

Topic (ii): Dissemination and client relations

# A STRATEGY FOR CONTROLLED AND SECURE ACCESS TO MICRODATA AT STATISTICS NETHERLANDS

**Invited Paper prepared by Olav ten Bosch and Frans Hoeve, Statistics Netherlands**

## Summary

**Abstract:** This paper describes the strategy developed at Statistics Netherlands to offer trusted researchers from Dutch universities and (governmental) research institutes access to well-documented microdata. It highlights the general policies for microdata access at Statistics Netherlands and it describes the actual microdata services being developed. It describes some of the infrastructural, organizational and security issues that come about when developing such a microdata dissemination strategy and finally it describes how facilities for microdata access could be integrated with other facilities for dissemination of official statistics, such as a statistical output database.

**Key words**: statistical information systems, microdata research, remote access, authentication, statistical disclosure control, statistical metadata, StatLine

## I. Introduction

1. During the last decade Statistics Netherlands observed an increasing demand for policy-related information. Policymakers need performance indicators to decide on the effectiveness of new regulations. In many cases, these detailed and very specific information requests cannot be answered from regular statistics and ad hoc statistical analyses on microdata available at Statistics Netherlands, possibly linked with microdata available at other institutes, have to be performed.

2. In addition to the growing number of policy-related information requests, the strategic focus shift from survey-based statistics to register-based statistics, as was mentioned earlier [1], created new possibilities for research institutes. This extended the opportunities to perform advanced research on data collections combined from different sources. Researchers showed their interest to perform additional research on the microdata sets that have been compiled by Statistics Netherlands for its own use from these surveys and administrative sources. In many cases researchers want to bring in their own datasets to be linked with the datasets of Statistics Netherlands to obtain a very specific dataset tailored towards their research interest. Of course, this type of research can only be performed under strict conditions of confidentially.

3. Dutch legislation makes it possible for Statistics Netherlands to meet the increasing demand for research on microdata. This paper describes the strategy developed at Statistics Netherlands to offer trusted researchers from Dutch universities and (governmental) research institutes access to well-documented microdata. It describes the actual microdata services being developed both from an organizational as well as from a technical point of view. It describes some of the infrastructural, organizational and security issues that come about when developing such a microdata dissemination strategy, the resulting microdata services that have been realized so far and the directions for extending these services in future.

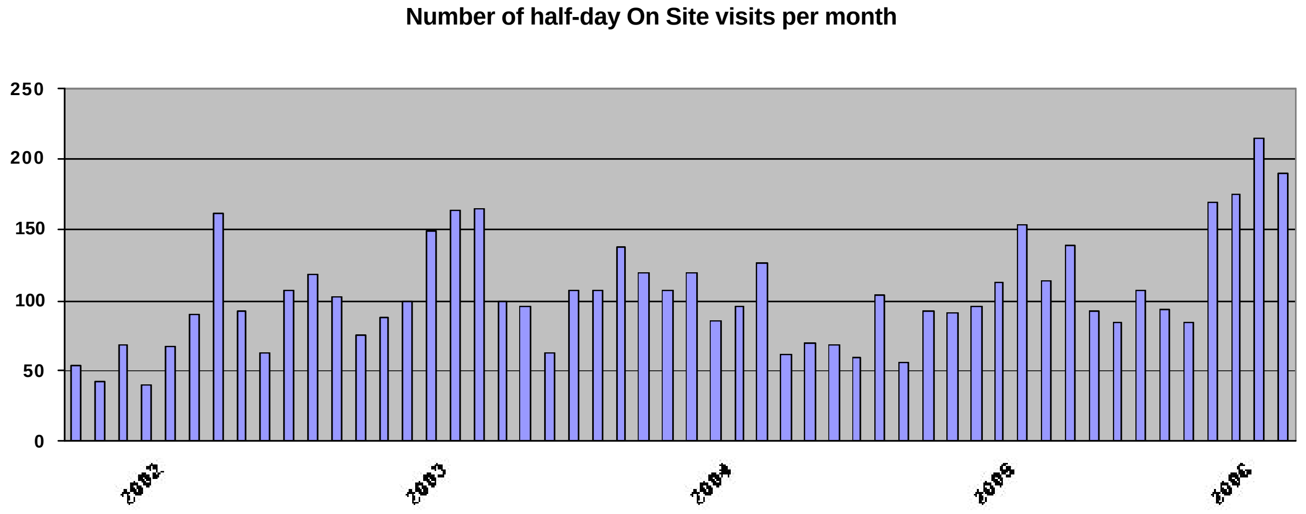
## II. One Strategy for all Microdata Services

3. As mentioned above, Statistics Netherlands has seen an increasing demand for dedicated analyses to be performed from policy-makers and from the research community. These analyses cannot be answered from regular statistics disseminated via the main website of Statistics Netherlands and require access to the underlying microdata. Although for reasons of privacy Statistics Netherlands cannot disseminate microdata directly, it does offer research institutes other ways to access microdata. Broadly speaking, these *microdata services* conform to the concepts and core principles for microdata access as described in [2]. In addition, the main strategy of Statistics Netherlands for microdata access is to offer a useful *mix* of services that together provide controlled and secure access to microdata and also provide a powerful and efficient environment for researchers to perform their task.

4. An important characteristic of the microdata services of Statistics Netherlands is that external researchers are allowed to work on microdata only under secure and confidential conditions. First of all, the services are open only to researchers of trusted institutes as specified by Dutch law or of institutes that have special permission to access microdata by approval of the Central Commission for Statistics of Statistics Netherlands. Second, each individual researcher has to sign a confidentiality statement and take notice of the rules for statistical disclosure control of Statistics Netherlands. Third, direct identifiers within microdata sets are removed or replaced by surrogate identifiers. This is done in such a way that linking different de-identified datasets is still possible. And finally, all results of microdata analyses are checked for confidentiality.

5. The first type of microdata service offered by Statistics Netherlands is the *On Site facility*. Researcher may work within one of the premises of Statistics Netherlands on a dedicated infrastructure. For reasons of security, this infrastructure is physically disconnected from the production environment of Statistics Netherlands and visitors only have access to the microdata

sets they need for their specific research. The environment is optimized for use by external researchers not familiar with the specific internal statistical tools and systems of Statistics Netherlands. The environment therefore offers popular tools such as Spss, Stata, Gauss and Ox. Sas may be used at additional costs. The results of their analyses are sent by email to their institutes, after a statistical disclosure check. This check can only be performed if the researcher provides enough information about the actual analyses performed. The figure below gives an indication of the use of the On Site facility (in terms of half-day visits per month) for the last few years. It shows that the actual use of the facility is still growing.



6. The second type of microdata service is called *remote execution*. Using this service, researchers may send in scripts to be executed on well-defined sets of microdata. They get their results back by email, but only after the usual disclosure check has been performed. Obviously, this can only work if the researcher knows the syntax and also the semantics of the microdata to work on. He may therefore request a synthetic sample of the dataset beforehand. The actual use of this facility during the last two years is limited. However, we expect remote execution to become more popular in the near future. Researchers who have worked on certain datasets before may know the microdata so well that they can easily do new analyses by remote execution.

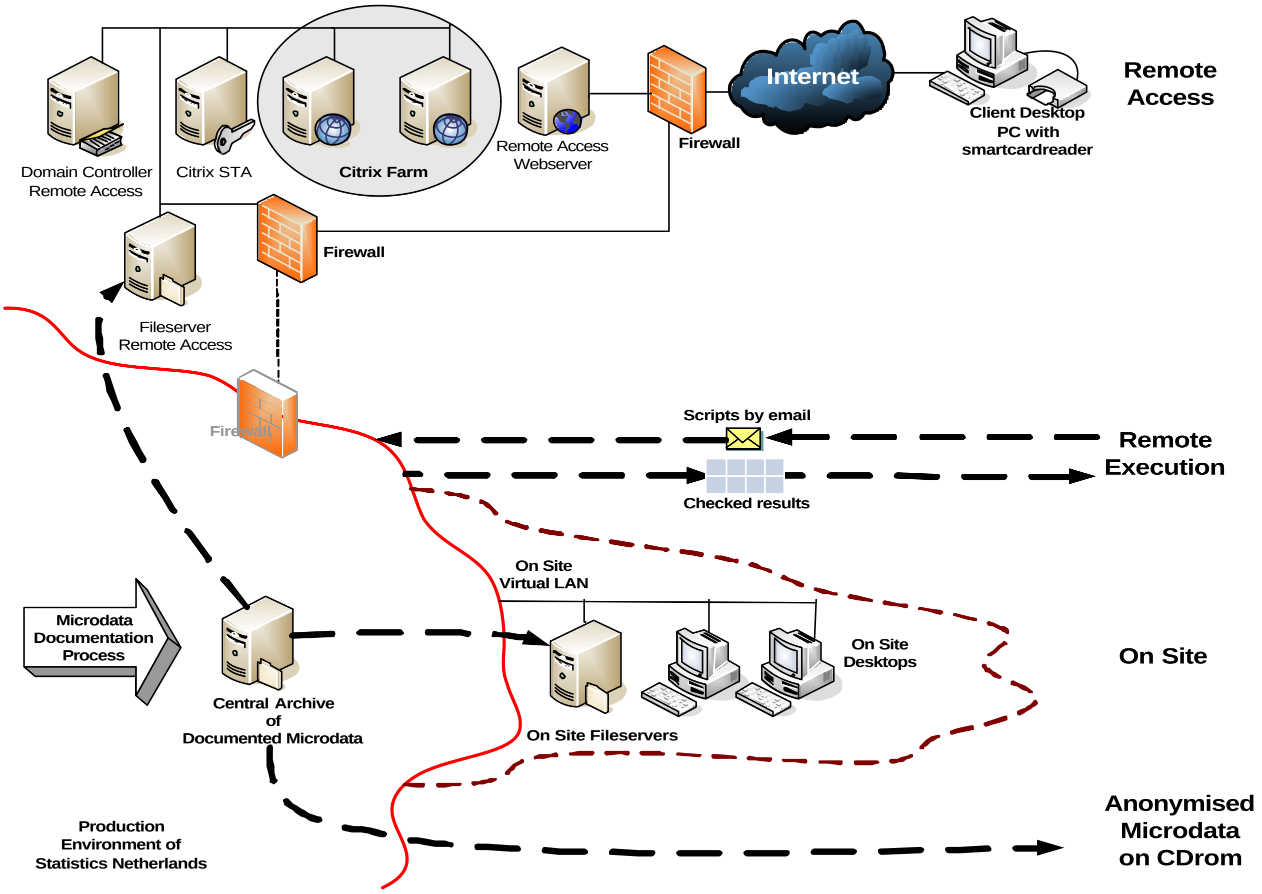
7. From mid-2005 a *remote access* facility is in pilot phase, which makes it possible for researchers to analyse microdata through a secure connection from workstations in their own institute. This facility is much alike the facility which is in use in Denmark [3], with the exception of the authentication method. From a technical point of view, the remote access facility is built up of a number of microdata servers, Citrix servers and a web server [4]. These servers are behind firewalls and disconnected from the production infrastructure of Statistics Netherlands. Applications can be launched from the web interface to run on microdata that has been imported to the (shared) workspace on the remote access data server beforehand. Authentication is implemented using biometric identification (fingerprints) and PKI certificates. Users may disconnect while their jobs keep running on the remote access server. If they log on again, they are automatically reconnected to these jobs. Ideally, the identity of the remote access would be rechecked during longer sessions, but this extra security check could not be implemented yet by our technical partner.

8. Researchers may work with the same statistical tools that are provided on the On Site environment, but an essential difference is that these tools do not run on their own workstation but on a server of Statistics Netherlands. Also, they cannot download the data to their own workstation, they can only “see” the data via the statistical tools they use. Of course, a researcher could in principle take a photograph of the screen or use optical character recognition (OCR) software to reconstruct the microdata on its own workstation. However, we make it clear in the regulations that this type of misuse is not allowed.

9. Checking results for disclosure is a labour intensive task, which cannot be automated easily, since it requires knowledge of the research field and an intelligent interpretation of the research results. This aspect needs special attention for the remote access service, since the number of remote users may grow heavily. The solution we are heading toward is to provide the means to write research reports via the remote access infrastructure, so that intermediate research results need no longer be checked. Instead, researchers may write a (draft) report which can be checked for disclosure. We expect this approach to keep the remote access facility scalable.

10. The last type of microdata service is the distribution of anonymised microdata files on CD-Rom via the Dutch Data Archiving and Networked Services (DANS) organization in the Netherlands. These microdata sets are completely safe in the sense that it should not be possible to extract any privacy sensitive information from it. This implies that the actual microdata sets disseminated via this channel may not express the same level of detail compared to the datasets disseminated via the other microdata services, which are more secure because of the extra disclosure control on research results.

11. The figure below gives a simplified view of the technical infrastructure and some of the data flows involved with the microdata services of Statistics Netherlands.



## III. Organization of Microdata Services

12. For historical reasons microdata services used to be spread across the organization of Statistics Netherlands, based on the type of microdata that was offered. Obviously, this conforms to the way Statistics Netherlands is organized internally, but on the other hand it is not very user friendly to a researcher that is not familiar with the internal structure of Statistics Netherlands. Therefore, recently all services for microdata research, from business statistics to social statistics, have been concentrated into one organizational unit, the *Centre for Policy-Related Statistics*. This organizational unit is the first entry point for questions related to microdata access. However, to answer these questions the centre works closely together with the internal experts on different microdata sets that are located in the statistical divisions themselves.

13. The microdata researchers may work on is documented especially for use by external researchers. This includes a detailed description of the variables, the quality of the data and information about the source of the data. In addition, it contains references to the results Statistics Netherlands published with respect to this particular dataset. This documentation standard was developed specifically to annotate microdata from social statistics, but it is currently being generalized to document microdata on business statistics as well. In addition, international metadata standards such as DDI [5] and SDMX [6] are taken into account when refining this standard in future. In all cases, researchers may bring in data from their own

institute to be combined with data from Statistics Netherlands as far as this is no risk from the viewpoint of statistical disclosure. Obviously, the combined data sets may only be used for their own specific research.

14. Statistics Netherlands has the policy that all output is available to everyone free of charge. There is no reason *not* to apply this principle to access to microdata. However, access to microdata differs from access to aggregated statistics in the fact that it requires additional work. Therefore, the pricing policy of Statistics Netherlands is that, although we do not charge for the microdata itself, the costs for using the microdata services are charged. This is inevitable, since the use of microdata services requires additional resources from Statistics Netherlands for using technical facilities, for creating additional documentation, for advice on the use of the microdata and for checking the results of researchers on disclosure risks.

## IV. Future Directions

15. In future, facilities for secure access to microdata could be integrated more closely with the dissemination facilities for aggregated statistics. For example, everyone can browse all information in the statistical output database *StatLine*, and in addition authorized researchers can “zoom in” to the underlying microdata sets that were used to create these results (top down). Ideally, microdata researchers working on a specific data set could easily retrieve the official statistics that were published based on these and other datasets (bottom up). Of course, researchers are only allowed to perform this type of operation if they have explicit permission to do so and when their identity is checked using biometric authentication. Unauthorized users cannot browse to the underlying microdata, but they do have access to the metadata of this microdata.

16. First of all, this would make Statistics Netherlands a more transparent organization to its customers. Second, this makes it possible for researchers to browse the metadata of the microdata archives for data that might be interesting for their research area. As a fist step in this direction, Statistics Netherlands decided to make the documentation of well-documented microdata sets available on the Internet and integrate them as much as possible with the thematic entrance to aggregated statistics on the website. This is only a first step to increased integration of micro and aggregated data in the statistical dissemination process. Even more powerful services, such as calculating statistics at run-time based on the interest of the user, may be achievable in future.

## References

[1] Olav ten Bosch, *New Developments in Information Processing at Statistics Netherlands,* Joint ECE/Eurostat/OECD meeting on the management of statistical information systems (MSIS), Geneva February 2003.

[2] Dennis Trewin, *Managing Statistical Confidentiality and Microdata Access – Draft Principles and Guidelines of Good Practice,* Work session on statistical data confidentiality, Geneva November 2005.

[3] Lars Borchsenius, *New developments in the Danish system for access to microdata*, Work session on statistical data confidentiality, Geneva November 2005.

[4] Anco Hundepool and Peter-Paul de Wolf, *OnSite@Home: Remote Access at Statistics Netherlands*, Work session on statistical data confidentiality, Geneva November 2005.

[5] The Data Documentation Initiative (DDI), http://www.icpsr.umich.edu/DDI

[6] Statistical Data and Metadata Exchange (SDMX), version 2.0, http://www.sdmx.org

-----