

Controlled and Secure Access to Micro data

Jan Mol, Frans Hoeve, Olav ten Bosch, 19-1-2009

1. Introduction

There has been an increasing demand for policy-related information in recent years, especially in the field of performance indicators to assess the effectiveness of new regulations. As the information required to answer these specific requests cannot usually be derived directly from regular statistics, specific statistical analyses have to be performed on micro-data. For scientific purposes, too, there is a growing demand for access to micro-data for analysis purposes.

Pushed by efforts to reduce the administrative burden generated by surveys in combination with the increase in the number of available administrative sources, there has been a large-scale shift from survey-based to register-based statistics. This has expanded opportunities to perform advanced research on data collections, especially where it is possible to link different datasets together.

Researchers have been showing an interest in performing additional research on the micro-data sets compiled by Statistics Netherlands for its own use from these surveys and administrative sources. Many researchers want to link their own datasets to those of Statistics Netherlands to obtain a very specific dataset tailored to their research topic.

This paper describes in short the strategy developed at Statistics Netherlands to offer trustworthy researchers from Dutch universities and (government) research institutes access to well-documented micro-data. It describes the main micro-data services from both an organisational and a technical point of view, and touches on some of the infrastructural, organisational and security issues.

Of course, this type of research can only be performed under strict conditions of confidentiality and has to comply with legal regulations. In this respect, Dutch legislation makes it possible for Statistics Netherlands to meet the increasing demand for research on micro-data.

2. Strategy for micro-data services

As mentioned above, there has been an increasing demand for performing dedicated analyses from policy-makers and from the research community. These analyses require access to the underlying micro-data. Although - for confidentiality reasons -

Statistics Netherlands may not disseminate micro-data directly, it does offer research institutes other means of access. The main strategy of Statistics Netherlands is to create a safe environment where researchers can access and analyse micro-data and where Statistics Netherlands retains control of all input to and output from this environment. This strategy focuses on three aspects: safe people, safe places and safe data. In this respect it is important to realise that the strategy is based on the twin principles of adequate security precautions taken by Statistics Netherlands, together with an explicit responsibility of the researcher.

3. Safe people

The micro-data services are available only to researchers of trustworthy institutes as specified by Dutch law, or of institutes that have special permission to access micro-data, subject to approval of the Central Commission for Statistics of Statistics Netherlands. To obtain this permission, the most important requirements are that their main activity is research, and that they make their publications available to the general public.

Individual researchers using the services are further obliged to sign a confidentiality statement. This confidentiality statement is also co-signed by their institute.

Before starting their work, researchers are made aware of Statistics Netherlands' rules with regard to statistical disclosure control. Without knowledge of these rules they cannot be expected to fulfil their obligation of confidentiality.

4. Safe places

To provide researchers with a safe environment to access micro-data, Statistics Netherlands has created a dedicated IT infrastructure. From a technical point of view, the infrastructure is built up of a number of micro-data servers, Citrix servers and a web server. These servers are protected by firewalls and are not connected with the production infrastructure of Statistics Netherlands. Applications can be launched from the web interface to run on micro-data that have been imported to the data server beforehand by Statistics Netherlands. The data themselves, as well as any application used to analyse them, remain on the servers of Statistics Netherlands. There are no printing, downloading or e-mail facilities within this environment. The user can see the data on his screen and see the results of his analysis, but these are only images sent to the screen. He cannot 'touch' the data themselves.

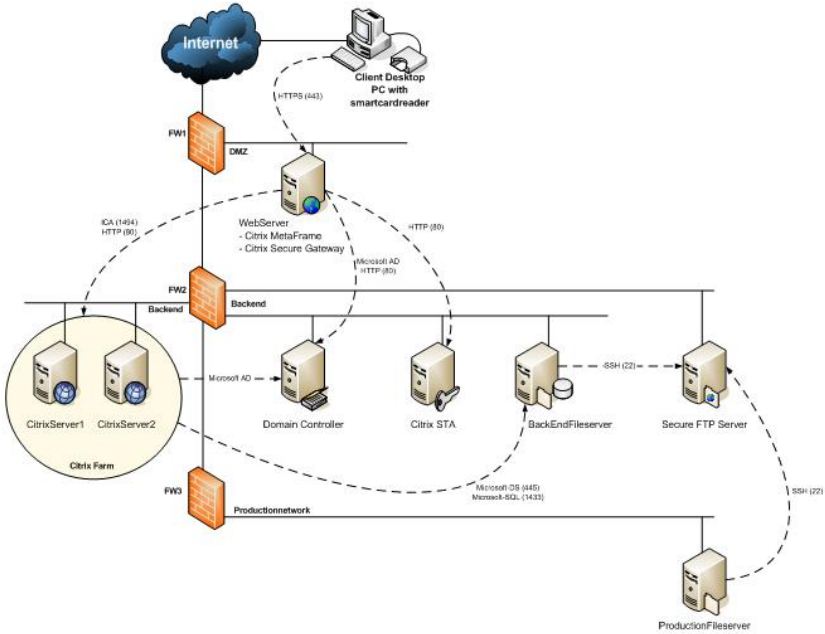
The environment is optimised for external researchers who are not familiar with the specific internal statistical tools and systems of Statistics Netherlands. The environment therefore offers popular tools such as SPSS, Stata, Gauss and Ox. SAS may be used at additional costs. For reasons of security, users have a personal account that only provides access to the micro-data sets they need for their specific research. The results of their analyses are sent by e-mail to their institute, after a

(manual) check on statistical disclosure. This check can only be performed if the researcher provides enough information about the actual analyses performed.

There are two ways to access this environment, *On Site* and by *Remote Access*. On-site access is situated in the premises of Statistics Netherlands. Both offices of Statistics Netherlands have a separate room where a total of 14 computers are set up with access to this environment through a username/password combination. Users can rent an on-site computer for a certain period.

The remote access option gives users access to this environment through a secure connection from a work-station in their own institute. Authentication is implemented using biometric identification (fingerprints). Users may disconnect while their jobs continue to run on the server at Statistics Netherlands. When they log on again, they are automatically reconnected to these jobs. The identity of the remote access user is rechecked during longer sessions

The figure below gives a simplified view of the technical infrastructure and some of the data flows involved with the micro data access facility of Statistics Netherlands. The on-site access is not explicitly shown in this figure, but it is simply a direct connection between an on-site computer at Statistics Netherlands and the web server.



5. Safe data

To prevent easy data identification, direct identifiers in micro-data sets are removed or replaced by surrogate identifiers. This process is performed by a small group of

staff who are entrusted with the keys to perform the translation between identifier and surrogate identifier. This is done in such a way that linking different de-identified datasets is still possible.

Checking results for confidentiality is a labour-intensive task, which cannot be automated easily, as it requires knowledge of the research field and an intelligent interpretation of the research results. This aspect requires special attention, as the number of users is growing rapidly, especially for the remote access facility. In order to limit the number of times a researcher has to withdraw output from the secure environment, he is provided with adequate 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.

In addition to the security of the micro-datasets, Statistics Netherlands pays special attention to the usability of the datasets. Therefore, all micro-datasets that researchers work on are documented in a specific way 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 publishes with respect to this particular dataset. This documentation standard was developed specifically to annotate social statistics micro-data, but it is currently being generalised to document business statistics micro-data as well. In addition, international metadata standards such as DDI and SDMX will be taken into account when refining this standard in the future. In all cases, researchers may introduce data from their own institute to combine them with data from Statistics Netherlands, as long as this does not give a risk of statistical disclosure. Obviously, the combined data sets may only be used for their own specific research.

6. Financial aspects

All services for micro-data research, from business statistics to social statistics, have been concentrated into one organisational unit, the *Centre for Policy-Related Statistics*. This organisational unit is the first entry point for questions related to micro-data access. However, to answer these questions the centre works closely together with the internal experts on different micro-data sets which are located in the statistical divisions themselves.

On principle, Statistics Netherlands makes all output available to everyone free of charge. Although there is no reason *not* to apply this principle to micro-data, such access differs from access to aggregated statistics in the respect that it requires additional work. Therefore, the pricing policy of Statistics Netherlands is that, although we do not charge for the micro-data themselves, researchers must pay for the micro-data services. This is inevitable, as the use of micro-data services requires additional resources from Statistics Netherlands for technical facilities, additional documentation, advice on the use of the micro-data, and checks of the results of

researchers on disclosure risks.

7. Future developments.

Statistics Netherlands would like to develop its micro-data facilities into a central node in the micro-data research landscape.

To do this, we are exploring possibilities to host data from other parties in our safe environment. In this way, researchers can access data from different organisations in the same environment, thus opening up new possibilities for linking datasets. At present, a first such pilot project with the Dutch National Central Bank is underway. Considering possibilities of setting up European networks to make cross-border access to micro-data possible is on the horizon. Statistics Netherlands is currently participating in several European projects aimed towards this goal.