OnSite@Home: Remote Access at Statistics Netherlands

Anco Hundepool[1] and Peter-Paul de Wolf[2]

[1] Department of Methods and Informatics, Statistics Netherlands, P.O. Box 4000, 2270 JM Voorburg, The Netherlands, e-mail: ahnl@cbs.nl
[2] Department of Methods and Informatics, Statistics Netherlands, P.O. Box 4000, 2270 JM Voorburg, The Netherlands, e-mail: pwof@cbs.nl

Abstract. In this paper we discuss a pilot project concerning a remote access facility at Statistics Netherlands. We describe some aspects of the technical implementation as well as the functional implementation. Moreover, we will discuss some tentative first experiences of external users.

Keywords. Statistical Disclosure Control, Remote Access, Citrix.

1 Introduction

Statistics Netherlands has a longstanding tradition of releasing safe microdata to researchers. This dates back to the beginning of the nineties of the previous century. The microdata files were made available to researchers at universities under a strict contract. These files were protected against statistical disclosure using a specific set of disclosure control rules that were defined based on the technological circumstances of those years. For many years these files could satisfy the research needs of the universities. The researchers could analyse the microdata files on their own computers.

However, the level of detail in these microdata files made it impossible for certain researchers to perform serious analyses. The restrictions of Statistical Disclosure Control (SDC), enforced by a national Law on Statistics Netherlands, did not allow more detailed microdata files to be made available to researchers outside the premises of Statistics Netherlands. The law demands that the use of and the results from analyses based on detailed microdata files should be under strict control of Statistics Netherlands.

To deal with this situation, the option to work OnSite, i.e., at the premises of Statistics Netherlands, was introduced: the detailed microdata files were made available to selected researchers in a controlled setting. The selected researchers could perform their desired analyses, but the results thereof were checked by Statistics Netherlands' staff for possible disclosure risk, before the researchers were allowed to bring the results outside the controlled setting. This means that only 'final' output (output to be taken home by the researcher) will be checked, whereas intermediate output can be examined by the researcher, but cannot be taken home.

The OnSite facility has proven to be very successful. For quite some time, many researchers have been using the facility. From time to time 5 to 6 researchers were working at the OnSite facility simultaneously.
A major drawback of this facility is that the researchers have to travel to the office of Statistics Netherlands, in order to be able to do their analyses. Even in a small country like the Netherlands this proved to be inefficient in many situations. Moreover, Statistics Netherlands has to organise specially equipped offices for the OnSite researchers.
As more and more facilities became available to use safe internet connections, the question has risen whether an equivalent of the OnSite facility could be build over the internet. This has led to the pilot project OnSite@Home. Only recently we have started a life test of this system as a pilot project with one partner at the University of Tilburg.

In section 2 we will describe the OnSite@Home facility in more detail, both from a functional and a technical point of view. The pilot with the partner at the University of Tilburg started very recently, but we will present some preliminary experiences in section 3. We will conclude in section 4 with some conclusions and remarks on the use of the facility as implemented for the pilot project.

2 OnSite@Home

In this section we will give a description of the Remote Access facility that is used in the pilot OnSite@Home. We will make a distinction between the technical and the functional aspects of the facility.

2.1 Functional aspects

The main idea is that the OnSite@Home facility should resemble the 'traditional' OnSite situation as much as possible, concerning confidentiality aspects. Moreover, it should resemble the look and feel of the OnSite facility without the aspect of having to travel to the premises of Statistics Netherlands. I.e., the following aspects have to be taken into account:

1. Only authorized users should be able to make use of this facility,
2. microdata should remain at Statistics Netherlands,
3. desired output of analyses should be checked on confidentiality,
4. legal measures have to be taken when allowing access.

2.1.1 Only authorised users are allowed

At the OnSite facility access of authorised users only is ensured because researchers cannot leave nor enter the premises of Statistics Netherlands unaccompanied. Moreover, only a selected group of researchers working at universities and similar institutes is allowed to make use of this facility.

The OnSite@Home facility is making use of biometric identification, to ensure that the researcher who is trying to connect to the facility is indeed the intended person. Whenever the researcher wants to access the facility, he will be identified by his fingerprint. For the pilot this is only checked at the start of each session, whereas in the future it is the intention to check the fingerprint at random times during the session as well. For more information about the process of logging on to the facility, see 2.2.1.
Obviously, again only a selected group of researchers will be given permission to make use of this facility.

2.1.2 Microdata remain at Statistics Netherlands

The network that is used by the 'traditional' OnSite facility is not connected to the production network. Moreover, the computers that the researcher can use are such that no removable media can be used (no floppy drive, no USB ports) and no internet connection is possible (no email, no surfing, no ftp, etc.). This means that the microdata used by the researcher can only be accessed using an OnSite computer at the premises of Statistics Netherlands and that the researcher cannot take a copy of the data to his institute. He is able to view the (intermediate) results of his analyses on the screen, but he is not able to send those results to his institute by email or otherwise. Moreover, he is not allowed to take a printout of the results to his institute either, without having it checked by a member of Statistics Netherlands' staff.
The network that is used by the OnSite@Home facility is separate from the production network as well (for the technical implementation, see 2.2). Moreover, the connection that a researcher makes with this facility is a terminal connection: he can only see on his screen what is running on a computer at Statistics Netherlands. He is not able to print or download any of the results of his analyses to his own computer at the institute where he is working. This ensures that the microdata and the intermediate results remain at Statistics Netherlands.

2.1.3 Checking output

Whenever a researcher wants to take output from a session at the 'traditional' OnSite facility to his own institute, the output first needs to be checked by Statistics Netherlands' staff for confidentiality. Only then he will either be allowed to take a printout with him, or the results will be sent to him by email.
A researcher that makes use of the OnSite@Home facility is not able to download or print his results either. Whenever he wants to have the results on his own computer at the institute, the results need to be checked by Statistics Netherlands' staff for confidentiality. The results will then be sent to him by email. For the technical implementation, see 2.2.3.

2.1.4 Legal measures

Both in case of the 'traditional' OnSite and the new OnSite@Home facility, legal measures are taken to prevent misuse of the microdata. To that end, a contract will be

signed by the institute where the researcher is working. Moreover, a statement of secrecy is signed by the researcher as well as the institute he works for.

2.2 Technical aspects

In Fig. 1 a simplified network representation of the OnSite@Home facility is given. An important fact is that there are three hardware firewalls involved denoted by FW1 through FW3, controlling the connectivity between the 'outside' world, the OnSite@Home facility and the production network of Statistics Netherlands.

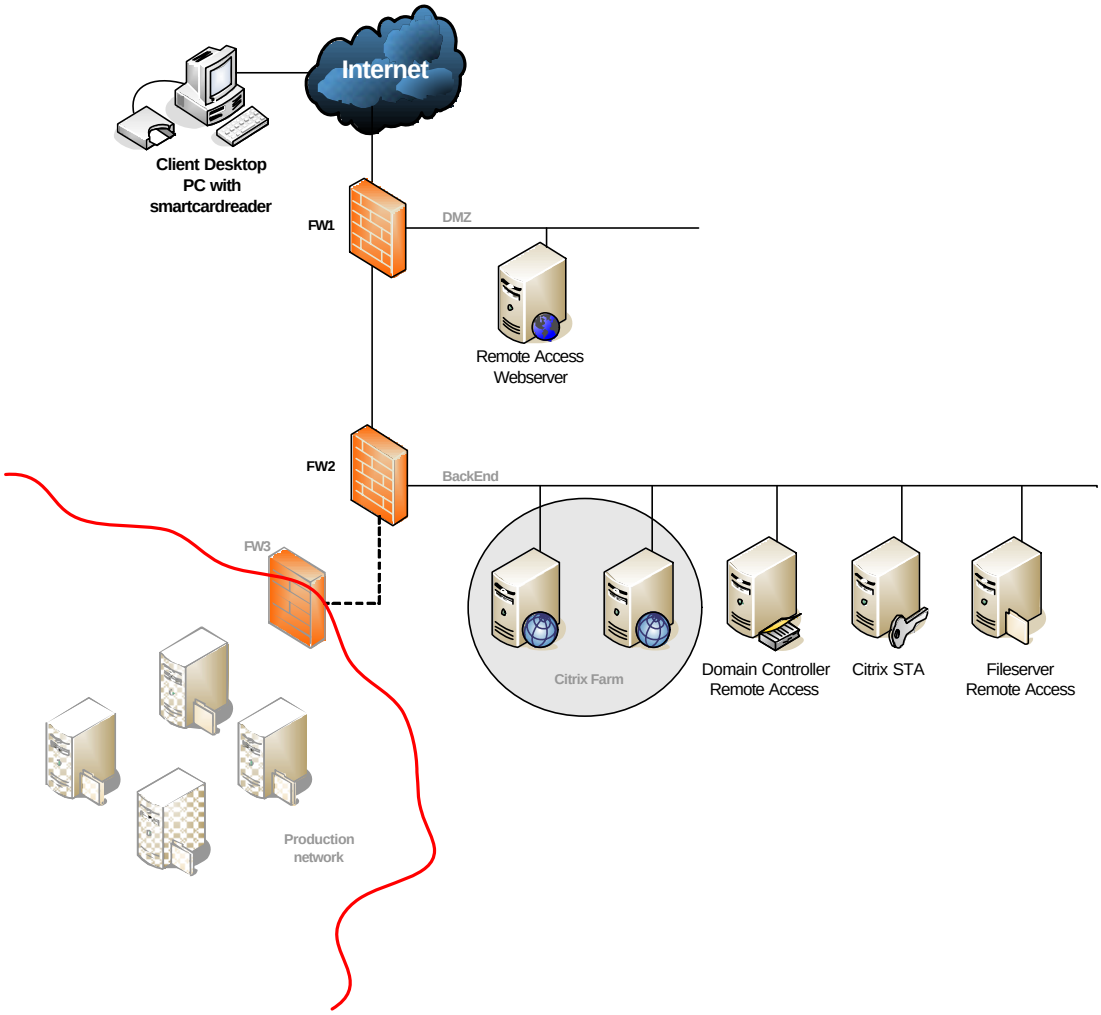

Fig. 1 Simplified network representation of the OnSite@Home facility

The servers in the Citrix farm and the Fileserver Remote Access, are separated from the other servers on the BackEnd using several different VLANs (virtual local area networks). Moreover, the different firewalls are configured such that only certain al-

lowed types of connection are possible between certain specific servers. E.g., the Remote Access Webserver is allowed to initiate some type of connection with a server in the Citrix Farm, but no server from the Citrix Farm is allowed to initiate any connection with the Remote Access Webserver.

The three firewalls effectively guard three parts of the complete network of Statistics Netherlands. The first firewall (FW1) controls the access of the 'outside world' to the demilitarized zone (DMZ). The second firewall (FW2) is in between the DMZ and the backend, where several 'intermediate services' are situated, like the Citrix part of the OnSite@Home facility and the e-mail servers for Statistics Netherlands (not displayed in Fig. 1.). The third and final firewall (FW3) separates the backend with the actual production network of Statistics Netherlands. This way it is virtually impossible to directly connect from an external computer to the production network.

2.2.1 The process of setting up a session

To ensure that only authorised users are allowed to set up a connection with the OnSite@Home facility, biometric identification is used, in combination with PKI[1] certificates. An authorised user is given a smartcard with a personal certificate. He will have to import the public part of that certificate onto his computer; the private part of that certificate is stored on an encrypted section that can only be decrypted by presenting the fingerprint that is also stored on the smartcard. This means that the user will need a smartcard reader that can read the users' fingerprint as well. This reader will be provided by Statistics Netherlands.
Whenever the user wants to start a session, he will have to start the Microsoft Internet Explorer and type the https address of the OnSite@Home facility to try to initiate an SSL[2] connection. Since it is possible that multiple researchers will make use of the same physical computer at the institute to access the OnSite@Home facility, he will then be prompted to choose which certificate he wants to use. Obviously, he is only able to use his own certificate: the private part of his certificate is written on the encrypted section of his smartcard. He will then have to present his finger to the fingerprint/smartcard reader in which the smartcard is inserted. If his fingerprint matches the one on the smartcard, the private part of his personal certificate will be released and sent to the Webserver. This server will check the credentials of the user using the Domain Controller. If everything is correct, the user will be shown the login site of OnSite@Home, using Citrix MetaFrame (a Web Interface). Finally, he has to type in his username and password to enter the main page of the OnSite@Home facility. On that page he will see the applications that he is allowed to use. For an example screenshot of that page, see Fig. 2.

[1] PKI = Public Key Infrastructure
[2] SSL = Secure Sockets Layer

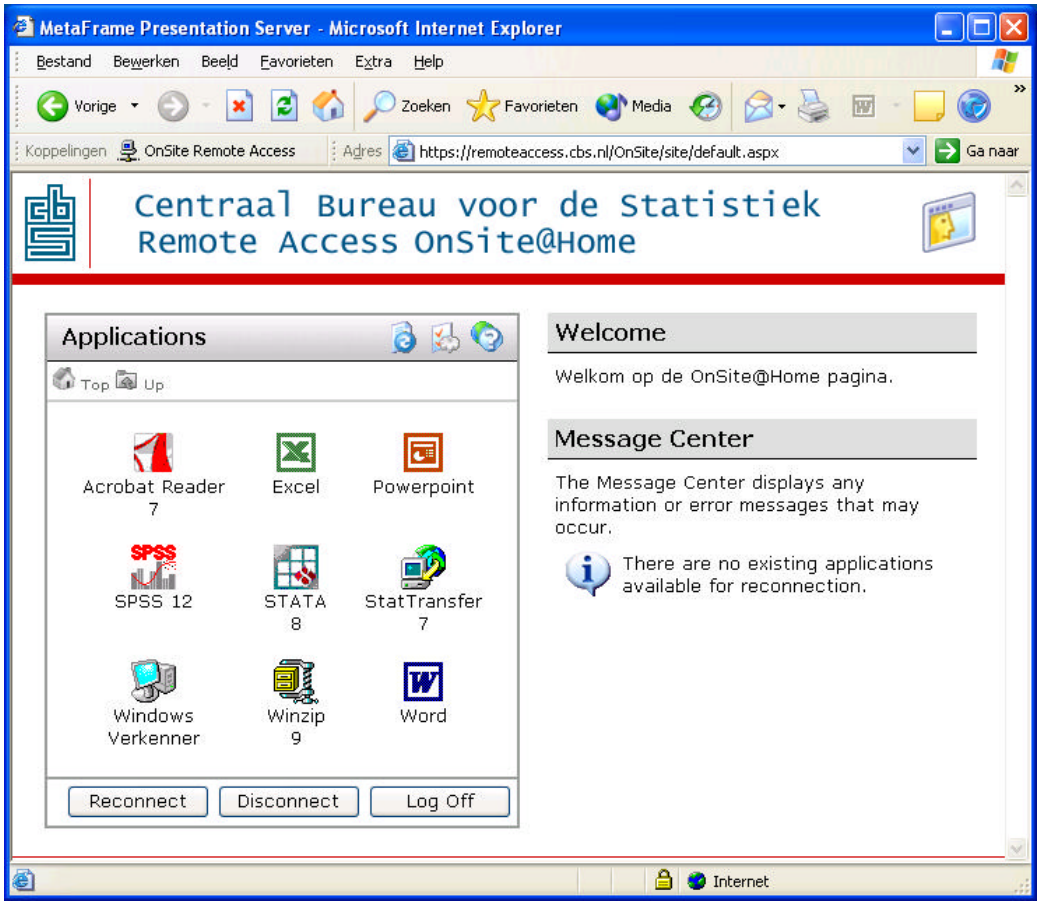

Fig. 2 Example screenshot of the main page of the OnSite@Home facility

2.2.2 Using an application

The researcher now has access to a number of applications and certain microdatafiles needed for his research. Moreover, he will have access to a working environment, in which he can store his intermediate results and/or files.

To start an application, SPSS say, he has to double-click the corresponding icon. The credentials of the user are then checked again and the Citrix STA (Secure Ticketing Authority) is asked to issue an ICA ticket[3] of limited lifetime. Using that ICA ticket, a secure connection will be established with one of the servers from the Citrix Farm. Then SPSS will be run on that server within the Citrix Farm. This connection is in effect a terminal connection: only screenshots of the connected server will be transmitted to the computer at the researchers' end.

[3] ICA = Independent Computing Architecture

2.2.3 Checking output for confidentiality

Using the OnSite@Home facility a researcher is able to perform his analyses interactively: he will constantly see what is happening on the terminal server in the Citrix Farm. I.e., he is able to see his intermediate results and to adjust his analyses accordingly. At some point, he would like to have some of his results on his own computer at the institute. Since it is impossible to print or download the results directly, the following procedure is set up.
The researcher places the results in a specific directory within his own working environment. A program, running on the Fileserver Remote Access, is constantly checking those directories and if anything is found in such a directory, that content will be placed on the secure ftp server. At the same time another program, running on a Fileserver on the production network, is constantly checking the secure ftp server for new content. If there is any on the secure ftp server, that program will move the content to the production network. In this way, there will never be a direct connection between the Fileserver Remote Access and the production network.
Once the output arrives on a Fileserver of the production network, a signal is given to specific Statistics Netherlands' staff. The output will then be checked for confidentiality and sent to the corresponding researcher by email within a half working day.
During the pilot, the check on the output for confidentiality is done by hand. Obviously, this is very labour-intensive. In the future, this should ideally be facilitated by some software. However, since the output of the results can be very diverse in format (SPSS, Stata, S-plus, SAS, etc.) the development of such a type of software is very difficult. Moreover, at Statistics Netherlands, no easily automated rules are available at the moment to decide whether or not general analysis' results breach confidentiality.

3 First experiences

The partner at the University of Tilburg (called Netspar) started to participate in this pilot project mid September 2005. At the premises of Netspar, two computers have been prepared for the use of the OnSite@Home facility, i.e., two smartcard/fingerprint readers were installed. Five researchers working at Netspar have been authorised to make use of the OnSite@Home facility and given their own personal smartcards. As far as the workload on the Citrixservers is concerned, this means that a maximum of two users from Netspar can be logged on to the system simultaneously, along with up to six members of Statistics Netherlands' staff (for testing purposes).

So far, no real problems have been encountered with the facility. Both the performance of the system and the look and feel resemble that of working on a state of the art workstation. I.e., it feels like working on your own computer.

4 Conclusions and remarks

The OnSite@Home facility seems to be a promising counterpart of the 'traditional' OnSite facility. Concerning confidentiality issues, both facilities appear to be comparable. The OnSite@Home facility is more flexible in allowing researchers to perform their analyses on microdata from a computer at their own desk, so they can work any time they want. Moreover, no travelling is needed whenever they want to perform additional research. On the other hand, the 'traditional' OnSite facility has the advantage that there will be no intervening colleagues while performing the research.

The technical implementation of the OnSite@Home facility tackles most of the confidentiality issues: the microdata remain at Statistics Netherlands, it is not possible to print or download any results and the final results will be checked for confidentiality before being released to the researcher.

Obviously, during the pilot we will monitor the experiences of the partner at the university of Tilburg. At the end of the pilot, an evaluation report will be written that can be used to further develop the facility and make it more generally available.

This facility can also be used to provide access to Microdata files Under Contract. Currently, those kind of microdata files are protected using statistical disclosure control methods as well as legal measures. These files are provided using CD-ROMs. Using the OnSite@Home facility, these files do not leave Statistics Netherlands, hence the dissemination of the microdata is much more under control. Moreover, a different level of statistical disclosure control might be possible.