Establishing Remote Access to Confidential German Micro Labor Market Data

Jörg Heining[1,2] and Stefan Bender[1]
[1] Institute for Employment Research (IAB), Nuremberg, Germany,
[2] Corresponding author: Jörg Heining, e-mail: joerg.heining@iab.de

Abstract

By implementing the 'Research Data Center – in Research Data Center' (RDC-in-RDC) approach, the Research Data Center (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB) in Nuremberg, Germany established for the first time remote access to confidential micro data in Germany. Remote data access systems, which allow researchers to access, evaluate and to see restricted micro data from their home desktop computer at any time, have not been implemented by a German Research Data Center (RDC) so far. Legal concerns, especially the problem of access control are reasons why German RDCs are not able to offer the research community these true remote access systems. The RDC-in-RDC approach overcomes these problems and may therefore be regarded as a first step towards true remote access in Germany. It may furthermore serve as a blue print for an intensified international data sharing. The basic idea is to allow remote access from designated institutions with comparable standards but locations other than Nuremberg. A thin client computer is used as an interface to establish a secure communication link to a server in Germany where the data are stored and processed. The connection to Germany is encrypted by using Citrix Software and a Citrix server. In a first step, remote access to FDZ data was established at four sites in Germany and one site in the US. In 2013, this remote access network will be expanded to more sites in Europe and the US. The RDC-in-RDC approach represents a change of paradigms in two respects. First, data access will be decentralized and data access is disseminated instead of data. Second, the dissemination of micro data is no longer restricted to national borders.

Keywords: Confidential Micro Data, Remote Data Access, International Data Sharing

1. Introduction

The Research Data Center of the German Federal Employment Agency at the Institute for Employment Research in Nuremberg, Germany provides researchers with access to confidential micro labor market data. The available data at FDZ on individuals, households and establishments come from several sources. Administrative data are obtained from the notification process of the social security system and the internal procedures of the Federal Employment Agency. The IAB also conducts its own surveys of households and establishments. The administrative and survey data can be analyzed separately, but all IAB surveys can also be linked to the respondent administrative records, resulting in, for example, linked-employer-and-employee data.

The FDZ applies a bundle of several strategies in order to ensure confidentiality and to prevent the disclosure of a single entity (see Hochfellner et al. (2012)). For example, these strategies include a conclusion of a use agreement with the researcher's institution, the stipulation and enforcement of contractual penalties in case of a violation of this use agreement, the anonymization of data, and the provision of restricted access ways to these data. In particular, the FDZ offers researchers three access modes:

- Scientific Use Files:
  Scientific Use Files are specially prepared data sets for off-site access. After the conclusion of a use agreement with FDZ, researchers can download a copy of the data to their local computer. Although this way of data access is very convenient for the researcher, it bears the disadvantage that SUFs are so called

factually anonymous data. Several anonymization strategies have been applied to SUF data resulting in a data product with minor analysis potential.

- Remote Execution
  Data with higher research potential so called weakly anonymized data may be accessed off-site via remote execution. Researchers submit codes to the FDZ which are processed with the data. The results are returned via email to the researcher after disclosure review.
- On-site Access
  Weakly anonymized data may also be accessed on-site at the FDZ in Nuremberg.

Especially to users of weakly anonymized data, on-site access is very important since this is the only access mode which allows to actually seeing the data. Although the FDZ provides documentation of the data and test data in order to prepare codes for remote execution, the disadvantage of not actually seeing the data cannot be compensated. Costly trips to access the data on-site at the FDZ in Nuremberg have been the only possibility for researchers if they wanted to actually "see" weakly anonymized data.

Several statistical agencies (see Bender and Heining (2011)) for an overview) have implemented so called remote data access systems which help to overcome the problem of costly visits to the on-site access facilities. In this context, remote data access is defined as the ability of a researcher to access and evaluate restricted micro data via a secure internet connection from his home desktop computer at any time. However, due to legal concerns remote data access systems have not been implemented by a German Research Data Center (RDC) so far. Especially the problem of access control, i.e. ensuring that the only approved persons access sensitive micro data remotely, is a reason why German RDCs have not been able to offer the research community true remote access to restricted data yet.

In 2010, FDZ introduced the Research Data Center-in-Research Data Center approach (RDC-in-RDC) which intends to overcome these legal concerns and to bring data access in Germany closer to the idea of remote data access. For the first time, researchers may access weakly anonymized FDZ data on-site at locations other than Nuremberg.

The remainder of the paper is organized as follows: section 2 introduces the general idea of RDC-in-RDC, section 3 describes the actual technical implementation. The current project status is depicted in section 4, an overview on challenges and future developments is given in section 5. Finally, section 6 concludes.

2. The RDC-in-RDC Approach

Nearly all RDCs, data enclaves or safe centers around the world share nearly the same standards with regard to the protection of confidentiality when providing access to sensitive micro data. The idea of RDC-in-RDC is to allow remote access to sensitive micro data from designated institutions with comparable standards as FDZ but locations other than Nuremberg. At these institutions (Guest-RDCs), a designated room will be equipped with specially configurated computers which allow establishing a secure communication link to a server in Nuremberg. All data processing is handled on this server. As a consequence, the sensitive data never leave the secure IT facilities of FDZ in Nuremberg. Given this set up, there is a clear division of work between the FDZ and the Guest-RDC. The FDZ handles the application process and concludes a use agreement with the researcher's institution. Once a valid use agreement exists, researchers contact the Guest-RDC in order to make an appointment for on-site access. The staff of the Guest-RDC verifies the identity of a researcher before he is provided access to the workstations supplied by FDZ. The staff of the Guest-RDC supervises the researcher during his session and enforces the rules of FDZ for on-site

access. Once the researcher is finished, the FDZ staff performs disclosure review and returns the cleared results to the researcher.

A more detailed description of RDC-in-RDC is given in Bender and Heining (2011).

3. Technical Implementation

The RDC-in-RDC approach is technically implemented by using a so called Citrix-thin client technology. Heining and Bender (2012) provide a detailed overview of this solution and other technical and organizational measures to ensure confidentiality within RDC-in-RDC. A thin client may be thought of as a reduced version of a regular desktop computer which only runs a minimum of software. All ports on this thin client are blocked preventing the connection of external optical devices, printers, or external storage media. Moreover, the application of security policies prevents the users to locally save files on the thin client. It is not possible to upload or to remove files to or from the thin client computer.

The purpose of the thin client computer is to serve as an interface for the user in order to establish a communication link to FDZ in Germany. All data processing is done on a server of the FDZ computing system. This computing environment is wholly contained and meets the security requirements of data that are designated as restricted access for confidentiality protection as stipulated by the German Law (in particular section 78a of the German Social Code, Book X and the Appendix to section 78a of the German Social Code, Book X).

The public internet is used for connecting the thin client at an external access point with the server in the FDZ computing system. The communication link is encrypted by using a Citrix web server and the Citrix Access Gateway software.

An important feature of the implemented technical solution is that two passwords are needed in order to provide a user access to FDZ micro data at the premises of a Guest-RDC. One password is only known to the data user and required for accessing the user's project directory in the FDZ computing system. The second password is kept by the supervisors of the Guest-RDCs. This password is needed for establishing the connection to the FDZ computing system. This construction ensures that in case of a theft of a thin client computer it is not possible to intrude into the FDZ computing system even when knowing one password.

4. Project Status

In 2011, the RDC-in-RDC approach was implemented at four sites in Germany. The RDCs of the Statistical Offices of the Länder (i.e., states) in Berlin, Bremen, Düsseldorf and Dresden provided a designated safe room for the RDC-in-RDC approach and have been equipped with thin client computers by FDZ. Moreover, a fifth access point was set up at the Institute for Social Research (ISR), University of Michigan in Ann Arbor, MI, USA.

The implemented Citrix-thin-client technology has proven itself as a stable system. System failures occurred very rarely and could usually be fixed within a couple of hours. From the beginning, all sites have been frequently accessed by data users making use of this new access way. In 2012, on-site access days at the Guest-RDCs (in total: 624) exceeded the utilization of the available workstations at the original Nuremberg site (505 access days).

Especially, the numbers for the Ann Arbor site are impressive (327 access days). Here, utilization was nearly 100 percent in 2012. FDZ identifies several reasons for this development. First, the data available at FDZ are comparatively easy to access. Similar data from the US are hardly accessible for researchers. Moreover, and in

contrast to the US, FDZ data are provided free of charge. However, the most important reason for the success of the Ann Arbor access point is equipped with permanent on-site staff by FDZ. The supervisor not only enforces the rules stipulated by German data protection legislation she also advises researchers during the application process and provides support when working with FDZ data.

5. Future Developments and Challenges

Expanding RDC-in-RDC

Given the success of the RDC-in-RDC approach, FDZ has been inquired by several national and international research institutions and data providers about the possibility of implementing the RDC-in-RDC at their premises, too. This already leads to the establishment of a sixth site at the University Applied Labour Studies of the Federal Employment Agency in Mannheim, Germany in April 2013. Moreover, the FDZ has agreed with the University of California at Berkeley and Cornell University at Ithaca, NY, USA on the opening of two additional access points in the US. Start of operation at these two new US sites is expected for mid 2013. In addition, FDZ has also started negotiations with Harvard University.

The RDC-in-RDC approach is also the favored mode of access within the Data without Boundaries project (DwB). For DwB, 29 statistical agencies, data archives, and research organizations from 12 European countries, joint forces to support equal and easy access to official microdata for the European research area. In the context of this project FDZ intends to open new sites at the University of Essex in Coulchester, UK and at the L'Institut national de la statistique et des études économiques (INSEE), Paris. Although not planed yet, additional sites in Europe and the US/Canada may be opened in the future.

Providing technical support

The operation of an international network of access points is a quite challenging especially with regard to providing fast IT support. A major problem is the time difference between Europe and the US.

Providing support on the data at Guest-RDCs outside Germany

As described above, data available at FDZ stem from administrative processes in social security and labor administration. In order to exploit the full analysis potential of the available data, data users need to have at least a basic knowledge on institutional and legal backgrounds governing the processes in social security and labor administration. For example, the functioning of the German unemployment insurance system may be well known to German researcher but may not necessarily be known to a researcher from the US. In order to support the special needs of international data users, not only documentation on the data is needed. Also institutional backgrounds governing the process of data collection need to be documented.

Currently, the FDZ staff in Ann Arbor is able to provide such information to data users in the US. However, funding for this position will expire by the end of September 2013. As a consequence, FDZ plans to integrate such kind of information to web based metadata system which allows international data users to easily access this kind of information.

6. Conclusions

By implementing the RDC-in-RDC approach the Research Data Center of the German Federal Employment Agency has established remote access to confidential micro labor

market data for the first time in Germany. In a first step, four external access points in Germany and one site in the US which provide on-site access to FDZ data have been established. Additional sites in Germany and abroad have already been opened or will start operation in the near future.

The experiences made so far are very positive. The utilization of the external access points is high, the applied technology runs stable. For the FDZ, the RDC-in-RDC approach represents a change of paradigms in two respects. First, data access is no longer centralized in Nuremberg. By conducting RDC-in-RDC, data access is disseminated instead of data. Second, the dissemination of micro data is no longer restricted to national borders. Thus, RDC-in-RDC may be a blue print for other data providers in order to establish transnational data access which may lead to an intensified international data sharing.

However, the development started by the RDC-in-RDC approach has not come to an end yet. The RDC-in-RDC approach is scalable in several dimensions thus providing a platform for further developments. So far, data of the FDZ are accessible at the premises of other data producers or institutions. A next natural step would be the establishment of the "opposite direction", i.e. the possibility of accessing confidential micro data of other, maybe even foreign, data producers at the FDZ. First talks about the implementation of an access network between several data producers in Germany already took place and a new project is about to be initiated.

References

Heining, J. (2010) "The Research Data Centre of the German Federal Employment Agency: data supply and demand between 2004 and 2009," *Zeitschrift für ArbeitsmarktForschung*, 42, no. 4, 337-350.

Bender, S., Heining, J. (2011) "The Research-Data-Centre in Research-Data-Centre approach: A first step towards decentralised international data sharing," *IASSIST Quarterly*, 35, no. 3, 10-16.

Heining, J., Bender, S. (2012) "Technical and organisational measures for remote access to the micro data of the Research Data Centre of the Federal Employment Agency," *FDZ-Methodenreport*, 08/2012, 14 p.

Hochfellner, D., Müller, D., Schmucker, A,. Roß, E., "Data Protection at the Research Data Centre," *FDZ-Methodenreport*, 06/2012, 27 p.