A New Access Mode to IAB's Scientific Use Files

Joerg Heining*
Institute for Employment Research (IAB), Nuremberg, Germany – joerg.heining@iab.de

Warren A. Brown
Cornell Institute for Social and Economic Research (CISER), Ithaca, NY, USA – warren.brown@cornell.edu

William C. Block
Cornell Institute for Social and Economic Research (CISER), Ithaca, NY, USA – block@cornell.edu

Stefan Bender
Institute for Employment Research (IAB), Nuremberg, Germany – stefan.bender@iab.de

Abstract

The Research Data Centre (FDZ) of the German Federal Employment Agency at the Institute of Employment Research (IAB) and the Cornell Institute for Social and Economic Research (CISER) teamed up to develop a new access mode to restricted micro data, so-called Scientific Use Files (SUF), specifically prepared for off-site access. This paper will introduce this new approach and demonstrate the benefits of this international data sharing initiative for academics and policy-makers.

Keywords: Restricted Micro Data; Off-Site Access; International Data Sharing.

1. Introduction

The Research Data Centre (FDZ) of the German Federal Employment Agency at the Institute of Employment Research (IAB) provides access to restricted micro data on the level of individuals/ households and establishments for academic purposes (see Heining 2010). FDZ was established in 2004 and is located in Nuremberg, Germany. Since 2011, it operates several remote sites both in Germany and the US. Data available at FDZ include social security records and data stemming from other administrative procedures of the labour administration, including worker level information on employment and daily wages, benefit receipt, as well as a rich set of socio-economic variables as of 1975 (for example the Sample of Integrated Labour Market Biographies (SIAB), see vom Berge, König, and Seth 2013). Moreover, FDZ provides access to linked data, where survey data are combined with administrative records resulting in unique (double) linked employer-employee data (see Heining, Klosterhuber, and Seth 2014).

Data files prepared by FDZ may be accessed free of charge and are available for researchers outside of Germany, too. In particular, FDZ has established three access channels. Approved researchers may access data on-site, i.e. either visit the Nuremberg site or one of the remote sites of FDZ in order to access and work with the approved data. In addition, FDZ also offers the possibility of remote execution. Researchers may send codes to be processed with the data to FDZ. FDZ reviews the output files in order to ensure the preservation of confidentiality and finally returns them to the researcher. Specifically prepared data files of off-site access, so-called Scientific Use Files (SUF), represent a third access mode at FDZ. After the conclusion of a use agreement with FDZ, researchers are provided with access to SUF which they may store and analyse on their local machines. Since SUF are also disseminated abroad, even researchers outside of Germany benefit from this access channel.

Although SUF are characterized by a comparable high degree of anonymity it is important to understand that SUF are still regarded as restricted data. This is expressed by the fact that SUF are only disseminated after the conclusion of a use agreement and therefore may not be simply available for download from the internet. Hence, when disseminating SUF to researchers, not only an adequate

preparation of the data to ensure confidentiality is important. Also the actual transmission strategy and how the restricted SUF data are protected locally by the researcher are challenging.
A strategy to tackle these challenges is represented by the joint project between FDZ and the Cornell Institute for Social and Economic Research (CISER) in Ithaca, NY, USA which aims to establish a new access mode to SUFs in the context of an international data sharing initiative. SUFs will be stored in the secure computing environment of the Cornell Restricted Access Data Center (CRADC) operated by CISER. In order to access SUFs, approved researchers will be provided with remote access to CRADC. Consequently, the risks associated with shipping and storing restricted data off-site are diminished.
The aim of this paper is to introduce this new approach and to discuss the benefits of this international data sharing initiative. Section 2 explains SUF in more detail while section 3 describes the dissemination strategies applied by FDZ so far and the problems involved. The new approach developed by FDZ and CISER is depicted in section 4. The benefits from this initiative are demonstrated in section 5. Finally, the section 6 concludes.

2. Scientific Use Files (SUF)

SUF are data products, specifically prepared by FDZ for off-site access (see Hochfellner, Müller and Schmucker 2012 or see vom Berge, Burghardt, Trenkle, and Seth 2013 for an example). They are characterized by a higher degree of anonymity than so-called weakly-anonymised data which are prepared by FDZ for on-site access. Weakly-anonymized data correspond to the original data but without any direct identifiers. In contrast, SUFs are classified as so-called factually anonymized micro data, a term defined in section 16, subsection 6 of German Statistical Act (Bundesstatistikgesetz). Data are regarded as factually anonymized if the information provided is substantially reduced and may only be de-anonymized with a disproportionate effort of time, cost and manpower. Several criteria for the preparation of SUF from a practical point of view are presented in Müller et al. 1991. The comparatively high degree of anonymity of SUF allows making such data available off-site and, moreover, that results based on SUFs need not to undergo disclosure review.
Access to SUF data is granted by FDZ in accordance with the regulations of section 282, subsection 7, Book 3 of the German Social Code (Sozialgesetzbuch). In general, access to SUF data is provided only for purposes of scientifically independent research. Researchers need to submit a proposal to FDZ with a non-technical description of the project that shows how the project is related to labour market research. FDZ evaluates the proposal along the dimensions of feasibility, appropriateness of the requested data and the risk to confidentiality by the project. Additionally, a data security plan which describes the technical and organization measures taken by the researcher to protect the SUF data when stored locally needs to be completed. In this security plan, researchers need to depict, for example, how the building and rooms in which the data are processed and analysed are protected, where the data will be stored (for example, server, local hard drive, or external drive), and by what technical measures and algorithms the computing system is protected.
After FDZ has approved the project, a use agreement between the researcher's institution and FDZ is concluded which defines the terms of data access and stipulates penalties in case of a breach or violation. It also contains both the (final versions of the) proposal submitted by the researcher and the data security plan. Once the use agreement is fully executed, FDZ transmits the data to the researcher.

3. Dissemination Strategies and Problems

In former days, FDZ transmitted SUF by sending CDs containing the data to approved researchers. Since CDs with restricted SUF data might get lost while in transit, FDZ changed the transmission process and now provides access to a secure exchange server from where approved researchers may download a copy of the data. Although the change in the transmission procedure towards a secure data exchange server solved the problem of data CDs getting lost while in transit by mail, this approach for the dissemination of SUF data is still characterized by some problems.

Access to SUFs is only granted for a limited time as specified in the use agreement. Once the use agreement has expired, researchers have to erase the SUF data from their systems and confirm the deletion to FDZ in written form. However, it is questionable whether the data have been actually removed. Copies of the data may still remain in the system due to back-ups or remain stored on a local machine which is exchanged in the course of the project. Moreover, the possibility of an intentionally violation of the use agreement by not deleting the data at the end of the project cannot be excluded. An additional problem is represented by the technical and organizational measures taken by the researcher to protect the locally stored data. As mentioned before, these measures are described by the researcher in a data security plan which becomes part of the use agreement. However, it is unclear whether the environment described by the researcher remains unchanged over the course of the project and continues to provide sufficient protection to the locally stored data. And, as before, the possibility of false statements or fraud exists.
A way of solving these problems would be audits of the facilities and systems by FDZ. All FDZ's use agreements contain respective clauses. However, given the number of current projects using SUF (more than 250) and given the fact that SUF data are also disseminated abroad, the idea of audits seems not to work in practice. In order to overcome these problems, FDZ teamed up with CISER and developed a new approach for disseminating SUF. The CRADC environment operated by CISER will host FDZ's SUF and will provide approved researchers with remote access to the data files.

4. SUF at CRADC

The Cornell Restricted Access Data Center (CRADC) was established by CISER in October 1999 as a pilot site sponsored by the National Science Foundation to provide secure access to confidential research data. Researchers can acquire, house, and use restricted data in CRADC's secure computing environment. It is a customized, state-of-the-art research facility to manage access restrictions required by data providers. CRADC staff works with researchers to tailor and implement security plans meeting provider requirements. CRADC'S secure computing environment consists of a Windows domain which is secured by a firewall and exceeds U.S. Defense Department C-2 standards for trusted computing environments. CRADC computing accounts are limited to those using restricted data for scientific research. Features of the CRADC environment include:

- Four 64-bit computing servers and a file server
- Access by Remote Desktop Connection or Terminal Services Client
- A domain controller employing user-based authentication
- Strictly enforced protocols for selecting and changing user passwords
- No connection to the outside world via FTP, e-mail, Web, print, or disk mapping facility

CRADC servers provide access to sophisticated statistical tools and support many software packages for data analysis and other tools for organizing researchers' work. All software is installed so that temporary files created by an application are saved in the data user's private disk space, not in areas where unauthorized users may have access.
In order to acquire access to SUF stored in CRADC, researchers need to submit a proposal describing their project to FDZ. The requirements for this proposal are identical as compared to the "traditional" dissemination procedure for SUF data. However, only a reduced version of the data security plan is necessary since the data are no longer stored on the local machine of the researcher. Once a use agreement is fully executed, FDZ asks CISER to activate an account for the specific project within CRADC. This account will be automatically disabled by CRADC after the use agreement has expired. Access to CRADC may be established either via a remote desktop connection or a terminal service client. Once logged to CRADC, the researchers may access several compute servers and a wide selection of software. File transfer to and from CRADC is possible, too. All programming, processing and data analyses are executed in the secure CRADC environment and researchers may remove output files from CRADC without disclosure review.

5. Benefits from the Joint Initiative between CISER and FDZ

The benefits of this new access mode to SUF data are obvious. Data are no longer stored on the local system of the researcher which prevents the possibility of accessing the data after the use agreement has expired. Moreover, this setting enables an easier and faster access procedure since researchers no longer have to ensure the security of the data by installing various technical and organizational measures.

However, the benefits of this joint initiative are not only restricted to an increase in data security and a facilitation of administrative procedures. CISER and FDZ teamed up on purpose in order to realize additional benefits. Both CISER and FDZ are leading experts in the production and dissemination of social science micro data. Each institution has highly trained staff, a respective organizational infrastructure, and deep experience in making complex and confidential data available to researchers. In the past, CISER and FDZ already successfully partnered for the implementation of a remote site of FDZ at Cornell University for on-site access to weakly-anonymised data within the scope of the so-called RDC-in-RDC approach (see Bender and Heining 2011 and Heining and Bender 2012). Given this joint history, it is expected that staff from both institutions will learn from each other and realize positive externalities when it comes to international data sharing and the development of new and innovative access modes to restricted data for academic purposes.

It is important to emphasize that this initiative to host SUF in CRADC also facilitates and enlarges access to data which have proven to be a unique resource for cutting edge research in Economics, the Social Sciences and Statistics. In general, Germany is third largest economy in the world with a workforce of 47 million people and an elaborate system of unemployment insurance and other social benefits. Therefore, the German labour market represents an interesting subject to study by itself. The relatively easy access to detailed longitudinal administrative data (since 1975) from FDZ provides an excellent basis to study the mechanisms governing labour markets and the incentive effects of labour market interventions. As a consequence, the insights gained from studies based on FDZ data may be directly applicable to other countries. Over the years, the high research potential of FDZ data and their applicability to a wide range of topics have led to numerous papers which have been published in several high ranked peer review journals, including the American Economic Review, The Quarterly Journal of Economics, the Journal of Labor Economics, among others. It is expected that this new access mode developed by CISER and FDZ will help to continue and support this development. It will allow academics to access SUF more easily which will foster the production of more papers and insights on the German labour market, thus providing policy-makers with more knowledge and a broader basis to base their decisions.

Easier access to data products of FDZ will also support the training and education of students and junior scientific staff. Average approval time until students can access SUF at CRADC is less than one month on average. Consequently, students may use FDZ data for class assignments as well as longer term projects such as master's theses and doctoral dissertations. FDZ is also open to group approvals so that students may use SUF for class. By working with SUF of FDZ, students and scholars learn how to work with administrative data and deal with problems. This creates valuable and transferable experiences/skills when they want to work with comparable data from different data producers/populations centres (for example, data from the U.S. Census Bureau, Longitudinal Employer-Household Dynamics (LEHD) data, etc.) for future research.

6. Conclusions

By storing SUF in the CRADC environment and providing researchers with remote access to this data, CISER and FDZ are currently developing a new access mode for restricted FDZ data which are specifically prepared for off-site access. This initiative will not only overcome some of the biggest problems when disseminating SUF, it will also help to realize positive externalities in data sharing. By establishing this new access mode, researchers and policy-makers will benefit from a wider

availability of FDZ data and will be able to acquire access to SUF more easily. This joint initiative between CISER and FDZ impressively demonstrates the benefits of international data sharing to scholars, policy-makers, and data providers.

References

vom Berge, Philipp; Burghardt, Anja; Trenkle, Simon; (2013): Sample of integrated labour market biographies * regional file 1975-2010 (SIAB-R 7510). (FDZ-Datenreport, 09/2013 (en)), Nürnberg, 73 S.

vom Berge, Philipp; König, Marion; Seth, Stefan (2013): Sample of Integrated Labour Market Biographies (SIAB) 1975-2010. (FDZ-Datenreport, 01/2013 (en)), Nürnberg, 65 S.

Bender, Stefan; Heining, Jörg (2011): The Research-Data-Centre in Research-Data-Centre approach: A first step towards decentralised international data sharing. In: IASSIST Quarterly, Vol. 35, No. 3, S. 10-16.

Heining, Jörg (2010): The Research Data Centre of the German Federal Employment Agency: data supply and demand between 2004 and 2009. In: Zeitschrift für ArbeitsmarktForschung, Jg. 42, H. 4, S. 337-350.

Heining, Jörg; Bender, Stefan (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 (en)), Nürnberg, 14 S.

Heining, Jörg; Klosterhuber, Wolfram; Seth, Stefan (2014): An overview on the Linked Employer-Employee Data of the Institute for Employment Research (IAB). In: Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften, Vol. 134, No. 1, S. 141-148.

Hochfellner, Daniela; Müller, Dana; Schmucker, Alexandra; Roß, Elisabeth (2012): Data protection at the Research Data Centre. (FDZ-Methodenreport, 06/2012 (en)), Nürnberg, 25 S.

Müller, Walter; Blien, Uwe; Knoche, Peter; Wirth, Heike. Die faktische Anonymität von Mikrodaten. Stuttgart: Metzler-Poeschel. 1991