Proceedings of Statistics Canada Symposium 2016
Growth in Statistical Information: Challenges and Benefits

A Modern Job Submission Application to Access IAB's Confidential Administrative and Survey Research Data

Johanna Eberle, Dana Müller, and Jörg Heining [1]

Abstract

The Institute for Employment Research (IAB) is the research unit of the German Federal Employment Agency. Via the Research Data Centre (FDZ) at the IAB, administrative and survey data on individuals and establishments are provided to researchers. In cooperation with the Institute for the Study of Labor (IZA), the FDZ has implemented the Job Submission Application (JoSuA) environment which enables researchers to submit jobs for remote data execution through a custom-built web interface. Moreover, two types of user-generated output files may be distinguished within the JoSuA environment which allows for faster and more efficient disclosure review services.

Key Words: Remote data execution; Micro data; Confidentiality.

1. Data and access modes at the FDZ

The Research Data Centre (FDZ) provides confidential micro data sets of the Institute for Employment Research (IAB) in a standardized way to (non-commercial) research institutions. Both administrative and survey data sets are offered. The data are sensitive micro data sets on individuals, households and establishments. The administrative data originate from employment notifications to social security providers and information on unemployment benefit receipt, registered job search and participation in labor market programs and training schemes as collected by the German Federal Employment Agency. The data collected from these processes are linked on the individual level so that a single individual may be tracked over time throughout its employment and unemployment history. All information is available on a daily basis and the earliest year of this data collection is 1975. IAB also conducts several comprehensive surveys covering different topics in employment research and related fields. These data are considered to be equally confidential as the administrative data. In addition, the FDZ offers linked employer-employee data as well as data sets linking administrative and survey information (Heining, 2010).

Data access is standardized and depends on the degree of anonymity of the data ranging from absolutely anonymous to highly detailed, so-called weakly anonymous, data[2]. With regard to the topic of this paper, only that type is described in more detail. Weakly anonymous data sets do not contain any direct identifiers such as name or address, but show a high risk of de-anonymization based on the amount and precision of the characteristics entailed. As a consequence, access to weakly anonymous data sets is only possible via on-site use or through remote data execution.

[1]Johanna Eberle, Institute for Employment Research, Regensburger Str. 100, D-90478 Nuremberg, johanna.eberle2@iab.de; Dana Müller, Institute for Employment Research, Regensburger Str. 100, D-90478 Nuremberg, dana.mueller@iab.de; Jörg Heining, Institute for Employment Research, Regensburger Str. 100, D-90478 Nuremberg, joerg.heining@iab.de

[2]Other types of data available at FDZ include Campus Files and so-called Scientific Use Files. Campus Use Files are absolutely anonymous and useful for academic teaching but not for valid substantial analysis. Campus Files can be downloaded by registered users agreeing on terms of use. In contrast, Scientific Use Files are factually anonymous data and are specifically prepared for off-site access. Despite a high degree of anonymity and the coarsening of variables and censoring of extreme values, Scientific Use Files contain more information than Campus Files and can therefore be used for meaningful empirical analysis. Scientific Use Files are transmitted to research institutions based on a data use agreement.

In order to be granted data access by means of on-site use and remote data processing, certain conditions need to be fulfilled in accordance to legal regulations (for more details see Hochfellner et al., 2014). The FDZ offers standardized request forms for the different modes of data access that clarify whether a research project complies with these conditions. After the data request has been accepted, the FDZ and the research institution conclude a specific data use agreement. The data can only be used for a specific project within a specific period as stated in the contract.

At the FDZ, weakly anonymized data sets are stored on file servers in an isolated network. For on-site use, visiting researchers have access to this network via local clients that do not have access to the internet or any other IAB network. Moreover, the FDZ provides additional workplaces for visiting researchers within this secure computing environment at different locations in Germany and abroad (Bender & Heining, 2011). Researchers can directly work with the weakly anonymous data but cannot download or transmit any of the results by themselves. Instead, results are submitted to researchers after disclosure review by the FDZ staff (for more details on data protection at the FDZ, see Hochfellner et al., 2012).

In contrast to on-site use, remote data execution means that researchers prepare their programs with artificial test data and submit these scripts to the FDZ. The programs are processed with statistical software and results are then provided to researchers after disclosure review by the FDZ staff. In the course of remote data execution, researchers do not have direct access to the original data.

Prior to 2015, remote data execution at the FDZ was only partly automated. Researchers sent their scripts via email to the FDZ and the FDZ staff manually copied these scripts to FDZ's isolated computing network. There, the scripts were processed with the data. Once the program runs were finished FDZ staff manually retrieved the results for disclosure review. However, over the last years, FDZ has experienced a substantial increase in the number of jobs submitted for remote execution, reaching a total of about 1,800 remote jobs in 2014. As a consequence, both managing remote data execution and providing disclosure review services for a continuously growing number of jobs became more and more challenging and time-consuming within the scope of this partly automated system. In order to deal with this situation, the FDZ decided to implement the JoSuA (Job Submission Application) environment, which will be described in the next paragraph.

2. Introduction of JoSuA (Job Submission Application) at the FDZ

The software bundle JoSuA was implemented at FDZ in April 2015. The software was developed and is maintained by the Institute for the Study of Labor (IZA). In order to meet the specific requirements of IAB's confidential data and to adapt the software to given organizational procedures at the FDZ, certain components of JoSuA were modified or extended with special features. Now, all relevant processes during remote data execution are handled within the JoSuA environment which executes most processes automatically. Only certain tasks such as disclosure review services still require actions by FDZ staff.

From a user's perspective, remote data execution is now processed exclusively through JoSuA's web interface (https://josua.iab.de). Researchers log into the web interface and upload their scripts. Beyond a standard web browser, no supplementary software is needed by the researcher. The jobs that were submitted to the webserver are subsequently forwarded to compute servers. These compute servers are in the secure and otherwise isolated network of the FDZ and have access to file servers holding sensitive data. Jobs are processed in this secure network with the statistical software package Stata. After the program run has finished, the output files are reviewed for any disclosure risk. Finally, all cleared output files are made available to researchers via the web interface.

As a counterpiece to the web interface, staff members of the FDZ use an operator's interface that allows them to conduct various administrative tasks such as overseeing running jobs, starting or ending processes, or managing user and project accounts. It also provides helpful functionality for disclosure review.

By May 2016, about 200 projects (usually including several researchers) actively used the JoSuA platform for remote data processing. The monthly average of jobs submitted is now about 700. Compared to less than 2000 jobs per year, the job count is now about four times the monthly job count of 2014.

3. Internal and External Output

Up to this point, the added value of the new job submission application is limited to a more advanced technical solution. The main innovation designed for JoSuA though is to distinguish between internal and external output, and to offer separate modes of job submission. The motivation is that only a small proportion of the output created during remote data processing and provided to researchers is actually used in a publication or a presentation. The vast majority of output is created for project-internal purposes only. Consequently, a large amount of output that is made available to researchers is only used to check the outcome of data preparation or to evaluate different types of analyses and models.

With this new distinction, researchers may now select among two modes of job submission: The first mode, "Presentation / Publication", is designed for those results that are actually to be used in a presentation or a publication. In this mode, output files are manually reviewed for any disclosure risk by scientific staff members of the FDZ and are subsequently made available as text files for download via the web interface. If researchers intend to publish results of their analyses of IAB data sets in a presentation or academic journal, book chapter, etc., they submit their job in this mode. Any output not fulfilling the criterion of absolute anonymity is not released (see Hochfellner et al., 2014).

The second mode, "Internal Use", can be selected during all preparatory steps of the research process as described above. In this mode, results can only be pre-viewed within JoSuA's web interface. For that purpose, the initial text output files are converted into image files and are then displayed within a viewing menu. A download of the image files is not possible. Since results are only available within JoSuA, manual disclosure control is replaced by script-based automated disclosure review. This is accomplished with an output anonymization script filtering the output based on regular expressions. Moreover, researchers are allowed to use the results of an "Internal Use" job only to prepare their analyses. Once a new job is submitted in "Publication / Presentation" mode, the results of past "Internal Use" jobs are inaccessible. Since output anonymization scripts are never all-encompassing, the remaining risks of disclosure in the view-only results are covered by a data use agreement concluded between the researching institution and the FDZ. In the view-only results window of the "Internal Use" mode, a watermark is shown in the background to remind users of the contractual commitment not to make screenshots or pictures of the results and not to make any internal results available to a third party outside the project's data use agreement.

One of the main advantages of this distinction is that only output required for a presentation or publication is given to the researcher. All other output generated in the course of the project remains within the secure computing environment of FDZ. This clearly increases data security and also minimizes the risk of disclosure. Furthermore, since jobs in "Internal Use" mode do not require manual disclosure control, results are faster available than in "Presentation / Publication" mode. Limiting the volume of output to be checked for any disclosure risks thus has an important efficiency aspect. Facing a growing number of research projects, putting the focus on output that is actually published guarantees that the volume of manual disclosure review is limited and still may be handled fast and efficiently.

So far, about 15 % of jobs are submitted by users in "Presentation / Publication" mode. The remaining 85 % of jobs are submitted in "Internal Use" mode. Disclosure control is performed by scientific staff and is very time-consuming. Averaged across 2015, disclosure control took about 16 minutes per job, with a range of one minute to 3 hours. In the first quarter of 2015, the total amount of time spent on disclosure control was about 120 hours. After implementing the distinction between internal and external output, the first quarter of 2016 now shows a reduced amount of 90 hours in total.

4. Conclusion

Although the introduction of JoSuA meant changes and adjustments for all sides involved, it can be regarded as a true success. The advantages for both the FDZ and the researchers of this new system for remote execution exceed the cost of adjustment by far.

One of the main advantages of JoSuA is that results from remote data execution are faster provided to researchers. Non-downloadable results are almost instantly available after the program run has finished. Moreover, JoSuA allows processing a larger amount of jobs for remote execution. As stated above, the monthly job count was multiplied by

four. By speeding up the research process and eliminating inefficiencies and frictions in the process, JoSuA therefore contributes fundamentally to benefit the scientific community.

Furthermore, since the amount of output transferred is now limited to those 15 percent of jobs that were submitted in "Presentation / Publication" mode, there is a reduced workload for manual disclosure review. Considering that at the same time, the total number of jobs has increased, the number of jobs to be manually reviewed is still reduced by about 25 % compared to the job count prior to the implementation of JoSuA.

References

Bender, S., and J. Heining (2011), "The Research-Data-Centre in Research-Data-Centre approach: A first step towards decentralised international data sharing", *FDZ-Methodenreport*, 07/2011 (en).

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, pp. 337-350.

Hochfellner, D., Müller, D., Schmucker, A., and E. Roß (2012), "Data protection at the Research Data Centre", *FDZ-Methodenreport*, 06/2012 (en).

Hochfellner, D., Müller, D., and A. Schmucker (2014), "Privacy in confidential administrative micro data – implementing statistical disclosure control in a secure computing environment", *Journal of empirical research on human research ethics*, 9, No. 5, pp. 8-15.