UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS
Expert Meeting on Statistical Data Confidentiality
26-28 September 2023, Wiesbaden

Remote Access for Scientific Use Files – a New Pathway for German Official Statistics Microdata Access

Hanna Brenzel (Research Data Centre of the Federal Statistical Office)
Katharina Cramer (Research Data Centre of the Statistical Offices of the Federal States)
Volker Güttgemanns (Research Data Centre of the Statistical Offices of the Federal States)
Marcel Mathes (Research Data Centre of the Statistical Offices of the Federal States)
Hanna.Brenzel@destatis.de

Abstract

The fundamental goal of the Research Data Centre of the Federal Statistical Office and the Research Data Centre of the Statistical Offices of the Federal States (RDC) is not only to provide access to official statistics microdata, but also to continuously improve and adapt the access to the changing needs of empirical science. In order to meet the broad range of needs of the empirically working scientific community, the RDC have offered different access paths since their founding, through which differently anonymised data products are made available. Now, the RDC come up with a new remote access prototype system including a new data product. All access paths differ both in terms of the anonymity degree of the provided microdata as well as in the access way of data provision. At first, existing and firmly established data access paths are outlined and their contractual and legal conditions explained. Subsequently, the newly installed remote access prototype and its features and requirements are presented. Provided that the ongoing evaluation phase turns out positive, this data access option will define one more way of data access operated regularly in its full version from 2024 onwards. The analysis potential of the data provided therein will classify between the scientific-use files transmitted to the scientific institutions and the data provided for on-site analysis at the RDC safe centres. This paper highlights various challenges, such as data protection requirements and legal framework conditions, which must be considered.

1 Introduction

With the establishment of the Research Data Centre (RDC) of the Federal Statistical Office in the fall of 2001 and with the RDC of the statistical offices of the Länder in April 2002, an important cornerstone and a central intersection was created between the scientific community and official statistics as data and information service provider.
Together, the RDCs offer the empirically working scientific community a coordinated range of data and services for the scientific use of high-quality microdata from official statistics.
Over time, however, expectations of the RDCs have evolved fundamentally, and stakeholders in politics and scientific communities have been pushing for substantial improvements in data access and data usage capabilities for some time.
Remote access represents an up to date and modern way of accessing data and is accordingly demanded by data users. The statistical offices of other European countries (e.g., the Netherlands, France or Finland) can be mentioned as reference benchmarks. They have created the legal and technical prerequisites to make their data available to researchers via remote access some time ago. Last but not least, a remote access system is currently being set up at European level by Eurostat.
On one hand, the establishment of a remote access system - with the investment in a connectable infrastructure - will advance the continuous development of the RDC. By catering towards the needs of the scientific community, the status of the RDC as a modern data provider will be consolidated. On the other hand, the currently complex and inefficient system of data access can be streamlined to a uniform and manageable system without limiting the flexibility of the users.

2 Status Quo

The RDC of the Federal Statistical Office, together with the RDC of the statistical offices of the Länder, offer access to more than 3,000 different data products for over 90 statistics for scientific use via different ways of access. They differ both in terms of the anonymity of the accessible data and in the type of data provision. Generally, the existing ways of data access can be divided into two categories, as figure 1 illustrates. In the case of the so-called "on-site access", the data remains in the secure areas of the statistical offices of the Federation and the Federal States. Since the RDCs can closely control the access to the data and provide output only after confidentiality check, the data are only weakly anonymized. With the "off-site access," on the other hand, users can work with the individual data at their own institutes. Since the output are not checked by the data centers, the individual data has to be more anonymized.
The category “off-site” includes the so-called Public Use Files (PUF), Campus Files (CF) and Scientific Use Files (SUF). “On-site” includes PC workplaces at the RDC, so called “safe centers” and remote execution (see the homepage of the RDC, https://www.forschungsdatenzentrum.de/en/access).

Safe centers exist in all locations of both RDC. These can be used by researchers to analyse microdata inside the safe premises of the statistical offices. As the individual data are already protected by the regulation of data access and the equipment of the PC workstation, formally anonymous microdata can be provided at the safe centers. Thus, a nationwide infrastructure in Germany is available for these data.
The safe centers are equipped with common statistical programs (Stata, R as well as partly SPSS and SAS) and are completely isolated from the outside. A separate PC workstation with internet connection is available for e-mail communication and internet searches.
In contrast to the safe centers, remote execution does not provide direct access to the microdata. Instead, data structure files are made available that resemble the original material with regard to structure and variable values, but do not permit any analyses in terms of content and do not hold any risk of exposing confidential information. Using these data sets, program codes can be prepared by the users using the statistical programs SPSS, SAS, Stata or R. These program codes are applied by staff of the statistical offices to analyse the original data. The data users receive the results of those analyses after the relevant confidentiality checks.

SUFs are standardized datasets created by the RDC for popular statistics. SUFs offer lower potential for analyses than on-site ways of access, but are designed to be suitable for a large proportion of scientific research projects. Due to the de facto anonymization of microdata, they may be used outside the protected premises of official statistics according to Sect. 16 para. 6 nos. 1 BStatG. Due to legal restrictions, SUF may only be used

by researchers who are employed by a research institution that is registered and located in Germany. The use of SUF may only take place in Germany. Until recently, the SUF were sent by DVD to the respective scientific institution with which user contract was concluded. Since June 2023, recent modernization measures now allow the SUF to be accessed directly via a download portal to the institution authorized to use the data.
In particular, on-site ways of access entail additional work for both data users and RDC staff. At the same time, the share of data uses via these access paths steadily increases over time compared to off-site uses. The development of a remote access system therefore pursues the goal of ensuring the technical connectivity to a modern and demand-oriented data provision for the scientific community. With this technology, the increased expectations of the research community for an up-to-date and modern data provision can be fulfilled in the long term. In addition, the remote access system holds potential for future innovation by reducing or substituting existing labor-intensive ways of access (reduction of on-site support, reduction of coordination of appointments with users, reduction of coordination and support of remote execution, etc.). Consequently, the scarce resources of the RDC could be invested more efficiently, for example in supporting additional data usage or further developing the data and service offers. At the same time, there is increased potential regarding data parsimony, as it is expected that this system will reduce the number of intermediate results per project that require confidentiality checks. Furthermore, the RDC aim to sustainably strengthen their leading role in the group of German RDC.

Figure 1: Ways of data access at the research data centres (RDC) of the statistical offices of the Federation and the Federal States

3 The Remote Access System

3.1 The technical structure

IT and data security play a crucial role in setting up the remote access system. The aim is to ensure that the remote access system is implemented in compliance with the law while maintaining the required IT security standards.

A virtual desktop infrastructure based on CITRIX was chosen as the IT-architecture. The system components set up are located in the so-called IDMZ (Internet Demilitarized Zone), in which procedures are operated that are to be accessible from the Internet. In the IDMZ, a distinction is made between three areas: Access Area (Pex), Application Area (Pin1) and Data Area (Pin2). These three areas are separated from each other by firewalls, which only allow approved communication between the neighboring areas within the application. A so-called transport encryption secures the communication path between the server and the client.

Two-factor authentication and IP whitelisting are implemented as additional IT security measures for the Citrix solution. Two-factor authentication means that, in addition to the user-specific work accounts protected by a personal password, a uniquely generated token must be used for each log-in. IP whitelisting allows only specific IP addresses to gain access to the remote access system. Prior to each authorized use, the IP address of the respective facility is allowed or added to the whitelist. This ensures that unauthorized IP addresses do not initially gain access to the system. This implements geoblocking as a technical measure as well as strengthening protection against possible (automated) attack attempts.

In addition, app protection is used to, among other things, prevent the user from taking screenshots of the data.

Remote system access is controlled on a per user basis by an access management system, only authorized users are granted access. Within the system, authorizations are limited to the extent required for data analysis. The creation of user-specific working accounts, which are managed centrally and secured by the user and access management, ensures that access is only possible to requested data. Each account is linked to a data folder in which user-specific official microdata are stored by RDC staff.

In addition to the technical measures, a number of technical and contractual-organizational measures are introduced to increase data protection. Before the data can be accessed, a user contract has to be concluded between the scientific institution and the responsible statistical office. It is contractually stipulated that up-to-

date software, operating system and virus protection are used on the client side when accessing the virtual desktop infrastructure. As well as, re-identification of individual cases is illicit. The RDC are legally bound to check all statistical results for statistical confidentiality that were created within the context of scientific projects based on provided microdata. This serves the protection of data according to section 16 (6) of the Federal Statistics Law (BStatG). Should individual cases be part of the output then they have to be blocked consistently across all results of a project. Data users who plan to re-identify individual cases are liable to prosecution and are expelled from further data uses.

In order to ensure that the system is tied to a specific location, its use is contractually established and sanctions are imposed in the event of violations. In addition, it is contractually stipulated that scientific institutions can be excluded from using the remote access system or from the possibility of carrying out further research projects via the RDC in the event of serious violations of the terms of use. In the event of a striking breach of contract, the scientific institutions can also be sanctioned with a penalty payment of up to EUR 20,000.

Figure 2: Technical infrastructure of the remote access system

3.2 Data material in the remote access system

Remote access to formally anonymized data is not feasible within the current legal framework. One possible way of implementation is to offer remote access for de facto anonymized data with slight modifications, as this would not require amendment of the law. In this case, the degree of data modification is of utmost relevance: If the level of anonymization is too high, the data offered will not meet the needs of the scientific community; if the level of data anonymization is too low, confidentiality can no longer be maintained. The degree of de facto anonymization therefore largely determines the benefits and coverage of the demand of the scientific community. In addition, the expected effects on the capacity of the RDC heavily depend on covering as many of the science community's projects as possible via the remote access system and, in particular, on reducing the costly uses of remote execution. However, this goal can only be achieved if significantly more data can be provided via remote access than via the current dissemination path via off-site SUF.

Microdata are described as “de facto anonymous” if it is not possible to completely rule out de-anonymization but assigning the information to the respective statistical unit “requires unreasonable effort in terms of time, cost and manpower” (Section 16 (6) of the Federal Statistics Act). According to the Federal Statistics Act, however, de facto anonymous data may only be used by scientific institutions and only to carry out scientific projects.

When creating de facto anonymity, the aim is to virtually eliminate the probability of correctly assigning data to respondents, while preserving the statistical information content as much as possible. Different anonymization methods can be used for this purpose. Common methods are information reduction (e.g. aggregation, class formation, censoring) and information modification (e.g. swapping). In order to determine de facto anonymity, the effort and benefit of deanonymization must be evaluated.

Factual anonymity thus does not completely exclude the possibility of re-identification, but puts its risk in a cost/benefit ratio. Costs for data users primarily include the consequences for actions in violation of the contract. Re-identification is strictly prohibited and punishable by fine or imprisonment (Section 203 StGB). In addition, consequences such as loss of reputation, loss of access to data of official statistics, etc., which threaten in the event of de-anonymization of the data, must also be considered by scientific users. This is because the users are obligated to maintain the anonymity of the data both by the formal obligation and the user agreement.

Factual anonymity therefore does not result solely from the remaining information content of the data, but is composed of a triad: 1) modification of the data material, 2) technical/organizational measures, and 3) contractual measures. Therefore, it also depends on the access condition, if a microdata set can be described as

de facto anonymous. Of crucial importance here is what additional knowledge is available and where the data access takes place. Depending on whether the microdata is used outside or inside the statistical offices, de facto anonymity can be achieved with more (off-site SUF) or less (on-site SUF) severe losses of information.

The de facto anonymity of microdata from official statistics is thus not a fixed quantity, but can be mapped along a continuum. In principle, it can be stated: The higher the technical and contractual measures, the fewer anonymization measures need to be taken and the higher the analysis potential of the data.

No technical measures are used for the previous off-site SUFs. Factual anonymity must therefore only be achieved from the two remaining measures: in addition to the contractual commitment and the commitment of the users, de facto anonymity is achieved by strongly anonymizing the data material itself. For this purpose, a statistics-specific anonymization concept is developed for each data material.

With the new remote SUF or on-site SUF, de facto anonymity can be achieved by significantly less modification of the data. This is justified by the high level of technical measures and the associated possibility to control the data access. In contrast to off-site SUF, the data is not passed on. It is solely possible to *view* the data via a virtual desktop (VDI environment). A so-called "transport encryption" secures the communication path between the server (sender) as well as the client (receiver). An exchange between the technical infrastructure of the data users and the data on the server of the official statistics or a download of the official data is thus technically impossible. Thus, unauthorized data linkage is impossible and the RDC has a high level of use control via log files. With regard to the risk of de-anonymization, data access via remote access therefore reduces many risks compared to the previous off-site SUFs.

3.3 The use of Remote Access

The remote access system, which is currently under construction, will be set up as a classic remote desktop version. As in the past, scientific institutions that are entitled to use the system in accordance with Section 16 BStatG have to apply for data access. If the application is approved, the researchers are then able to access the secure area within their scientific institution by using their own hardware. Within the secure area common statistical software such as RStudio and Stata is available. The major advantage compared to remote execution is that researchers can see the microdata and do not have to "blindly" program their syntaxes as before (see Figure 3). By working directly with and being able to view the data, it should be possible to significantly reduce the number of intermediate results previously generated via remote execution, thus minimizing a very labor-intensive process step in the RDC. The goal should be that only final outputs are checked for confidentiality by the RDC staff and will be released. This also supports the principle of data parsimony.

Figure 3: Remote Access at the RDC

Work on setting up such a system began in November 2021. The system is currently in the evaluation phase. On one hand, the technical implementation of the system is being tested and its resilience checked using penetration tests. On the other hand, the user-friendliness and the attractiveness of the data material provided is to be examined thoroughly. In a first step, only absolutely anonymous data material was made available via the system for a selected group of people. In a second step, off-site SUFs will then be made available to power users who have already completed a valid user application with the RDC. The third step will then be to test the redesigned on-site/remote SUF material. Since the system requires a redesign of all statistics-specific anonymization concepts, a gradual integration of the existing data products in the RDC is planned. The start will be made with the most requested data product, the microcensus. In order to be able to evaluate the operating grade of the system appropriately, DRG statistics will be offered as one of the first data products in the remote access system in addition to the microcensus. If the evaluation of the system is positive, other data products that are of high demand will follow.

4 BIBLIOGRAPHY

Brenzel, Hanna / Zwick, Markus. An information infrastructure has emerged in Germany – the Research Data Centre of the Federal Statistical Office. German version published in WISTA | 6 | 2022, p. 54 et seq.

Homepage of the Research Data Centre of the Federal Statistical Office and the Federal States
https://www.forschungsdatenzentrum.de/en