UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert Meeting on Statistical Data Confidentiality

26-28 September 2023, Wiesbaden

Remote access to European microdata

A. Bujnowska, F. Espelage, M. Stocchi (Eurostat – statistical office of the European Union)

Aleksandra.BUJNOWSKA@ec.europa.eu

Frank.ESPELAGE@ec.europa.eu

Marco.STOCCHI@ec.europa.eu

Abstract

In 2021, the European Statistical System Committee agreed to enable remote access to highly detailed European microdata: secure use files. Secure use files are microdata to which no further methods of statistical disclosure control have been applied. Access to secure use files is regulated by Commission Regulation (EU) No. 557/2013 and has traditionally been performed by researchers visiting the safe centre located at Eurostat premises in Luxembourg.

To implement remote access procedures, Eurostat organized consultations in order to discuss the legal, organisational and technical aspects of remote access to secure use files; successively, Eurostat developed the functional architecture of the remote access system in accordance with the relevant methodology for data preparation and research output checking. The information system (hereinafter “KIOSK”), eventually delivered by the Direction General for Informatics of the European Commission (EC), is composed of a central server farm endowed with remarkable scalability features and compliance to the security standards of the EC. It can be accessed from end-points hosted by accredited research entities of the Member States of the European Union and EFTA countries. KIOSK offers: i) continuous availability of data and metadata to researchers, whose project proposals were successfully approved; ii) software tools for data analysis; iii) tools suitable to data protection and statistical disclosure control (SDC). The system entered the production phase in January 2022. Besides deploying in KIOSK several software platforms and tools popular among the community of statisticians, Eurostat decided to invest in innovation, exploring possible solutions to longstanding data protection problems such as research output checking (ROC). Among the several different categories of ROCs, Eurostat opted for the automatic analysis of research output based on known theoretical SDC rules (principle-based output SDC), featuring partial automation: first automated output checks are applied, then the resulting output is finally evaluated by a specialist in the field of output checking. The ROC tool in use (named “ACRO”) has been designed in order to privilege simplicity of use and low training overhead, thus having a concise set of functional requirements. It implements automatic checks on the tabulation of descriptives (such as frequencies, percentiles, moments, maxima and minima), several popular estimators and statistical tests. ACRO is expected to work with primary suppression only, with aggregates recomputed post-suppression, as required. Secondary suppression was not considered as optimal SDC solution in the case of ROC, because of its error-proneness and higher likelihood of disclosure risk by differencing. Developed by a team of experts of the University of West England in the form of a proof of concept in STATA®, ACRO has been uploaded to a public Git repository to the benefit of research and statistical institutions, and it is currently installed in KIOSK. Eurostat is planning to extend the ACRO functionalities and, eventually, its support to several different open source scripting and programming languages.

1 Introduction

European microdata represent a precious and unique source for the international research community. As microdata, they consist of sets of records containing information on individual respondents or business entities. As European microdata, they are available with common structure and harmonised content for all EU countries, thereby allowing detailed analyses and comparisons at EU, country, and sometimes even regional level[1].

To protect the anonymity of respondents (persons, organisations), access to European microdata is regulated and restricted. Two types of confidential data for scientific purposes exist, namely secure use files and scientific use files.

Regulation (EU) No 557/2013[2] is the legal base for scientific access to European microdata. For secure use files, it provides for access within access facilities (1) at Eurostat or (2) accredited by Eurostat. In practice, until 2021, access to secure use files could only be performed by researchers visiting the safe centre located at Eurostat premises in Luxembourg. This paper informs on the project allowing remote access to European secure use files via accredited access points. It also identifies some challenges and suggests directions of future work.

2 European Statistical System and European statistics

Eurostat is the statistical office of the European Union. Its mission is to provide high-quality statistics and data on Europe. The data necessary for the performance of the activities of the EU are determined in the European statistical programme. These data are called European statistics.

The European Statistical System (ESS) was established by Regulation (EC) No 223/2009[3]. It is the partnership between Eurostat and the national statistical institutes (NSIs) and other national authorities (ONAs) responsible in each Member State for the development, production and dissemination of European statistics[4]. The ESS is governed by the ESS Committee bringing together the heads of the NSIs and Eurostat.

Regulation 223/2009 determines the main principles of development, production and dissemination of European statistics and establishes the governance of the ESS. The data necessary for individual domains are defined in subject matter regulations[5]. These regulations lay down the details of the data collection at the national level and the requirements for data transmission to Eurostat. A harmonised approach to data collection and processing allows for EU comparable data.

The subject matter regulations define as well in which format data have to be sent to Eurostat. Whenever microdata are transmitted, Eurostat may propose to add the respective microdata collection to the list of microdata available for scientific purposes.

A separate legal act – Regulation 557/2013 – establishes the conditions under which access to confidential data transmitted to Eurostat may be granted.

[1] Eurostat receives data from EU countries as well as from Iceland, Norway, Liechtenstein, Switzerland, and candidate countries.

[2] Commission Regulation (EU) No 557/2013 of 17 June 2013 implementing Regulation (EC) No 223/2009 of the European Parliament and of the Council on European Statistics as regards access to confidential data for scientific purposes.

[3] Regulation (EC) No 223/2009 of the European Parliament and of the Council of 11 March 2009 on European statistics.

[4] Most of the European legal acts (including the regulations mentioned in this paper) are also applicable to Iceland, Liechtenstein and Norway. Application in Switzerland is based on bilateral agreements between EU and Switzerland.

[5] Not all domains are governed by regulations; some data transmissions from statistical offices to Eurostat are voluntary.

3 Access to European microdata for scientific purposes

Regulation 557/2013 is the main legal framework establishing conditions of access to European microdata. Operational guidelines (Guidelines for the assessment of research entities, research proposals and access facilities – the Guidelines[6]) accompany the regulation and provide further details and explanations.
Amongst other things, the guidelines describe the process of preparation of datasets for research use. This process is performed in cooperation between Eurostat and NSIs. It includes decisions on the type of microdata for research, and on statistical disclosure control (SDC) methods applied on the data[7]. European microdata released for scientific purposes never contain any direct identifiers.
Confidential data for scientific purposes can be made available in two types, as:

- Scientific use files, or
- Secure use files.

Scientific use files (SUFs) are the most popular among researchers. These are partially anonymised European microdata available for download from a secure Eurostat platform. Researchers may use SUFs only at the premises of recognised research entities. The methods of partial anonymisation are agreed beforehand with the NSIs; typical methods are global recoding, top- and bottom-coding, micro aggregation etc.
One standard version of anonymised microdata is prepared for each data collection (like EU-SILC, LFS…). Usually, the same SDC methods are applied on subsequent versions and releases of the data. However, changes to the anonymisation approach may occur, e.g. in reply to user needs, or to take changes in the data structure or content into account. Such changes require consultations with NSIs and their agreement again.
Secure use files are European microdata to which no anonymisation / SDC methods are applied. These data are only available in the secure environment of Eurostat. At the moment, secure use files exist only for two business data collections (Community Innovation Survey (CIS) and Structure of Earnings Survey (SES)). All results produced by researchers using secure use files must be checked for confidentiality (output checking), and only validated output may leave the secure environment. The output checking rules are also agreed with NSIs.

4 Access to secure use files and development of remote access

Until 2021, access to secure use files could only take place in Eurostat's safe centre. During the COVID-19 pandemic, Eurostat's safe centre was closed, and the need to offer remote access to secure use files became a priority. In a general attempt to modernise access conditions, Eurostat had started the development of an IT remote access system some years before, so the required IT system (called KIOSK) was already available. The remaining legal, administrative, methodological and technical issues, however, had still to be agreed with NSIs in order to make this system available to users.

In the beginning of 2021, a dedicated task team bringing together representatives of NSIs was set up. For several months, Eurostat together with the task team members worked on legal, administrative, methodological and technical aspects of remote access. In October 2021, specific conditions to allow remote access to secure use files were proposed to and agreed by the ESS Committee. These conditions are included in the guidelines (see section 3). They are presented in the following section.

5 Remote access to European secure use files – general conditions

Remote access to secure use files may take place from accredited access points located in recognised research entities[8]. A research entity wishing to host an accredited access point must fulfil all conditions specified below:

[6] Guidelines for the assessment of research entities, research proposals and access facilities (https://ec.europa.eu/eurostat/documents/203647/771732/guidelines-assessment.pdf)

[7] Countries may opt out from inclusion of their data in confidential microdata for scientific purposes.

[8] Recognised research entities are entities eligible to request access to European microdata. These are usually universities, research institutes or research departments of other organisations. In order to be recognised, research entities must provide evidence that they fulfil the following conditions: (1) research as main objective of the entity; (2) scientific publications; (3) independence in formulating scientific conclusions; and (4) data safeguards in place. More details in the guidelines and on Eurostat website: https://ec.europa.eu/eurostat/web/microdata. The full list of research entities recognised by Eurostat is

1. Be recognised as a research entity by Eurostat;
2. Be located in one of the following countries[9]:
   - EU countries, Iceland, Norway and Liechtenstein (covered by Regulation 557/2013 and by the GDPR) and Switzerland (covered by Regulation 557/2013 and by the GDPR on the basis of a separate agreement), including EU institutions, bodies and agencies and international organisations based in these countries;
   - countries covered by adequacy decisions[10];
3. Fulfil the organisational requirements specified below and in the rules of use of access points;
4. Fulfil the technical requirements specified below and in the rules of use of access points.

5.1 Organisational requirements for access points

The research entity must designate **a person responsible for the access point**. This person must:

1. Ensure the use of the access point by authorised persons only: this means researchers employed by or linked with the research entity (see chapter 4.2.1 of the guidelines) with a validated research proposal mentioning the use of remote access; only persons authorised for the same project may be present in the access point at the same time;
2. Provide the services necessary for smooth and proper running of the access point (for example: run a system of access point reservation, collect statistics on access point use etc.);
3. Retain physical access logs for at least six months and make it available to the Commission when requested;
4. Ensure that the access point and the computers located in the access point comply with the technical requirements specified below and in the baseline security requirements (for access points and for computers) defined by Eurostat[11];
5. Ensure that the access point is used in accordance with the terms specified in the Rules of use of access points, confidentiality undertaking and terms of use of confidential data;
6. Follow up immediately any abnormality, security incident or violation of the rules; this includes informing Eurostat immediately.

5.2 Technical requirements for access points: access point setting

The access point must be a separate room or rooms (depending on the needs) located in the premises of a research entity. The access point shall be set in a way to reduce physical risks such as eavesdropping, unauthorised observation of activities and loss or theft (the access point cannot be the open, public space).
The door to the room must be lockable.
When not in use as an access point, the room might be used for other purposes (for example access to other sensitive data). These purposes must not be conflicting with the use as an access point.

5.3 Technical requirements for access points: computer setting

The computers in the access point must comply with the following requirements:

1. Have broadband internet connection;

also available on Eurostat website: https://ec.europa.eu/eurostat/documents/203647/771732/Recognised-research-entities.pdf.

[9] The accreditation of access points located in other countries is possible in principle, but under additional conditions and contractual clauses as used for transferring personal data to non-EU countries. This option will be developed in a later phase.

[10] https://ec.europa.eu/info/law/law-topic/data-protection/international-dimension-data-protection/adequacy-decisions_en (please note that Canada will be treated as third country, as the adequacy only concerns commercial organisations).

[11] The requirements will be published on the Eurostat website (https://ec.europa.eu/eurostat/web/microdata) so that research entities can align to them.

2. Have a dedicated fixed public IPv4 address. Only computers physically located in the access point room(s) should be able to use this IP address;
3. Have Firefox (minimum version 78.11.0esr) web-browser application installed;
4. Have remote access **to** the computer disabled (physical access on-site only);
5. Allow only the connection of screen, keyboard and mouse peripherals; the connection of external storage devices must be blocked;
6. Have all screen capture and sharing capabilities disabled;
7. Have the videoconferencing tools disabled.

6 Experiences and challenges identified so far

In 2022, Eurostat started a pilot implementation with JRC-ISPRA. Until mid-2023, only one further recognised research entity (University of Copenhagen) applied successfully for hosting an accredited access point.
Even if access to secure use files at Eurostat's safe centre had always been difficult (travel to Luxembourg, limited time in safe centre, host person at Eurostat required…), the new option of remote access to secure use files is not yet as popular as expected. Eurostat promoted it e.g. via a microdata access newsletter end 2022, but there are some issues preventing a bigger interest:
First, the organisational and technical requirements are quite strict. A particular technical difficulty seems to be the dedicated fixed public IPv4 address as it is not necessarily in line with the general IT rules and procedures of interested recognised research entities.
Second, the limited offer of secure use files (Community Innovation Survey (CIS) and Structure of Earnings Survey (SES)). Basically every presentation of the remote access option immediately triggered requests for more data available that way, especially EU-SILC and LFS.
Third, the limitation of access to researchers of the research entity hosting the accredited access point. Discussions with data archives revealed that hosting an access point would be very interesting for such organisations if they could offer remote access to researchers of other research entities.

7 Directions for further development of the service

7.1 Tool for Automatic Checking of Research Outputs (ACRO)

As mentioned above, results of the analysis of European secure use microdata must be controlled for statistical confidentiality. Researchers must get Eurostat approval for all statistics they would like to take away from the secure environment (no matter if access in Eurostat's safe centre or via remote access point).
In order to facilitate the work of output checkers, in 2019 Eurostat initiated the work on a tool doing such checks automatically. This led to development of ACRO: Automatic Checking of Research Output[12]. ACRO is a proof of concept developed in 2021. It assists researchers in their data analysis, identifying confidential results which would not pass the confidentiality check. The researchers can modify the data until they reach an adequate level of confidentiality. The researchers may do this dynamically, on their own, without involving output checkers, which is an important efficiency gain. The positive feedback received so far from both researchers and output checkers encourages Eurostat to further improve ACRO especially in the direction of making ACRO available for different platforms (currently ACRO works in STATA® only). Another path of development is enlarging the scope of functions currently dealt with by ACRO. At the moment, ACRO only works on few output types (tabulations, common estimators, medians, maxima/minima).

7.2 Projects to have more accredited access points and more datasets available

It is the objective of Eurostat to increase the number of accredited access points and to allow more researchers to work remotely with secure use files.

[12] To find out more about ACRO see Eurostat Statistical Working Paper "Automatic Checking of Research Outputs (ACRO): a tool for dynamic disclosure checks". The ACRO source code is published on https://github.com/eurostat/ACRO. See also proceedings of the 2021 Eurostat/UNECE work session on statistical confidentiality.

Eurostat will hence continue its efforts to promote the remote access service and will strive to make the information about the relevant requirements easily accessible and visible on Eurostat website[13].

It is also the intention of Eurostat to increase the offer of secure use files, but that is a longer-term project requiring the necessary resources. Inclusion of additional domains is only possible after

1. NSI agreement on secure use files for e.g. EU-LFS, EU-SILC…. (secure use files should offer 'significantly' more than existing scientific use files);
2. NSI agreement to allow access to the respective secure use files with their country data via remote access;
3. Organisation of output checking.

[13] The link to Eurostat microdata access website: https://ec.europa.eu/eurostat/web/microdata