Access to European Statistical System Microdata

Aleksandra Bujnowska

1 Introduction

Microdata play an essential role as a primary data source in the production of official statistics. In addition to their use for statistical purposes, the potential of microdata for policy and scientific purposes has been increasingly recognised over recent years. Their analysis being facilitated by technological developments, microdata are extremely valuable as they allow assessment of the underlying structure and causal links of the studied phenomena.

National statistical offices in the European Union (EU) Member States and Eurostat can make microdata available to users for research purposes. While practices to grant access to microdata at national level vary from one country to another, microdata held by Eurostat for all EU Member States (and in some cases European Free Trade Association (EFTA) countries) are provided to researchers according to a transparent approach, in line with applicable legislation.

This chapter focuses on the organisation of access to microdata produced by official statistics, and in particular by the European Statistical System. Sections 2 and 3 explain basic terms and concepts of microdata access. In Sect. 4 the elements of the generic microdata access system are presented. Section 5 then introduces the European microdata access system. Finally, Sect. 6 concludes with some indications on the way forward.

All views expressed are those of the author and not of the Commission (Eurostat).

A. Bujnowska (✉)

European Commission, Eurostat – Statistical Office of the European Union, Luxembourg City, Luxembourg

e-mail: aleksandra.bujnowska@ec.europa.eu

© The Author(s) 2019

N. Crato, P. Paruolo (eds.), *Data-Driven Policy Impact Evaluation*,
https://doi.org/10.1007/978-3-319-78461-8_6

Fig. 1 Statistical data available from Eurostat, NSAs and other sources

2 The European Statistical System and European Statistics

The **European Statistical System (ESS)** is a partnership between Eurostat and the national statistical institutes (NSIs) and other national authorities responsible in each Member State for the development, production and dissemination of European statistics. **National statistical authority (NSA) is a generic term for NSIs and other national data providers** (e.g. regional statistical offices, ministries providing administrative data, etc.); a list of NSAs is available on the [Eurostat website](#).¹

European official statistics are important for EU. They are produced and disseminated by Eurostat in partnership with NSAs. Usually, national official statistics are based on microdata, collected or accessed by NSAs. Microdata are then aggregated, transmitted to Eurostat and published. Where necessary for the production of European statistics, NSAs also transmit microdata to Eurostat (see Fig. 1). Whenever microdata are transmitted, Eurostat may consider granting access to these for scientific purposes. In this way, almost all microdata received by Eurostat are released for scientific purposes.

¹Path: Eurostat website/About Eurostat/Our partners/European statistical system.

3 Microdata Access Terms and Concepts

Microdata are a form of data where sets of records contain information on individual persons, households or business entities. Traditionally, statistical offices use microdata only to produce aggregated information such as tables. Publication of individual information (microdata) is generally not allowed because it may easily lead to identification of the data subject (person, household or business entity) and therefore to a breach of statistical confidentiality.

Statistical confidentiality is one of the fundamental principles of official statistics. It is the obligation of the statistical offices to protect confidential data.² In the context of European statistics, confidential data are data that allow the identification of statistical units (individual persons, households or business entities), thereby disclosing individual information. The statistical unit may be identified in the different forms of statistical output, e.g. the contribution of largest companies may be approximated in business statistics. To prevent this, statistical offices check each output from the point of view of statistical confidentiality. This check is called **statistical disclosure control (SDC)**.

The **SDC** methodology helps to identify confidential data in these various output forms and to hide such data, taking into account relationships between the data (e.g. additivity of the tables).

In general, official statistics are available in the form of tables where confidential data are not visible and the data are highly aggregated. But many statistical offices also make available their data in the form of microdata, namely as (see Fig. 2):

- Public-use files accessible to everybody (sometimes upon registration or licence signature)
- Confidential microdata files accessible to researchers satisfying specific access conditions

Confidential microdata files are invaluable for the research community as they allow deep analysis of relationships in the data, i.e. causalities, dependencies, convergences, etc. Microdata access systems were developed by statistical institutes to allow legitimate access to confidential data for scientific purposes.

²The ESS and, in a broader sense, official statistics are legally obliged to respect statistical confidentiality. The entities collecting data for purposes other than statistical ones (e.g. commercial, administrative or health purposes) fall into the scope of personal data protection legislation. Statistical confidentiality protection measures are stricter than those stemming from personal data protection legislation.

Fig. 2 Types of data made available by statistical offices

4 Elements of the Generic Microdata Access System

Microdata access systems define under which conditions access to confidential microdata can be granted for external persons, such as researchers. These conditions are normally outlined in legal acts. In the European Statistical System, access to microdata may be granted to researchers carrying out statistical analysis for scientific purposes.³

Microdata files may have different levels of detail. The more detailed the data, the easier it is to identify individuals. **Original statistical records** can be easily identifiable as they contain unique direct identifiers such as names, address, social security number or identification number (ID number). These confidential records with direct identifiers are available to the statistical offices only under strict confidentiality protocols.

Microdata without direct identifiers are called ‘**de-identified**’ or ‘**pseudonymised**’ **microdata** (if direct identifiers are replaced by pseudo-identifiers: unique codes replacing all direct identifiers). De-identified microdata with pseudo-identifiers are more and more important for the production of official statistics, as they allow linking data collected from different sources, thus fostering the use of, for example, administrative sources and derivation of further results on the basis of already collected data. Pseudo-identifiers also allow the creation of longitudinal files, following individuals over time. These microdata are still confidential, as the combination of some rare characteristics may lead to identification of unique statistical units.

De-identification is a subprocess of anonymisation. In general, anonymisation is the process of making the data anonymous. However, approaches to this process differ between countries. In some countries, making the data anonymous is defined as removal of names, i.e. de-identification. In the European law, **anonymisation** is defined as the process aiming at complete protection of microdata, such that

³Article 23 ‘Access to confidential data for scientific purposes’ of the Regulation (EC) No 223/2209 contains enabling clauses for access to ESS microdata.

the records are no longer identifiable (the records cannot be linked to any ‘real’ person, household or business entity). The different stages of microdata anonymisation/protection are (see Fig. 3):

- De-identification or pseudoanonymisation: process of removing direct identifiers (such as name, ID number and address) from the confidential data, and replacing them with pseudo-identifiers. Pseudo-identifiers can be used to link datasets.
- Partial anonymisation: application of a set of SDC methods to microdata in order to reduce the risk of identification of the statistical unit. Scientific-use files are the result of partial anonymisation.
- Complete anonymisation: application of SDC methods that completely eliminate the risk of identification of the statistical unit (directly or indirectly). Public-use files contain completely anonymised records.

Table 1 compares all basic types of microdata files and access conditions.

The terms **secure-use files** and **scientific-use files** are specific to the European microdata access system. In the EU countries, there exist similar files but with different names, e.g. scientific-use files are often called ‘microdata files for research’. The basic characteristics of these files remain the same:

- Secure-use files are files to which no further methods of statistical disclosure control have been applied. Researchers access these files in the secure environment provided by NSAs (local or remote access). The final results of the work of researchers are checked by NSAs to ensure that they do not reveal confidential data. Each output is checked separately.
- Scientific-use files are files to which methods of statistical disclosure control have been applied to reduce (not to eliminate!) the risk of identification to an appropriate level (partial anonymisation). Researchers have access to such files outside the controlled NSA environment. There are usually no ex post controls by NSAs; researchers need to follow the confidentiality instructions and are responsible for making the published results non-confidential.

Secure use files are the richest form of microdata for research. However, the services related to provision of access are usually expensive for statistical offices. This is because of infrastructure (dedicated environment for on-site or remote access) and operational costs related to output checking.

For statistical offices, scientific-use files seem to be more efficient in terms of cost-benefit ratio. For researchers, the advantage is that they can be used without having to travel to the premises of the statistical offices (or without logging in to a remote, secure system).

Scientific-use files may be standard or tailor made, i.e. adapted to the particular needs of the research project. The risk of a breach confidentiality is smaller if standard files are released than if specific files are produced on request. For researchers, however, the standard files are often not sufficiently detailed (e.g.

Fig. 3 Anonymisation processes and the resulting types of microdata files

Table 1 Characteristics of the different types of microdata

	Original microdata	Microdata for statistical purposes	Secure use files	Scientific-use files	Public-use files
Identification risk level	Extremely high	Very high	High	Low (reduced)	Eliminated
How the respondents can be identified	By direct identifiers	By combination of indirect identifiers (characteristics such as NUTS ^b level, size class, NACE ^c category)	By combination of indirect identifiers (characteristics such as NUTS ^b level, size class, NACE ^c category)	By combination of indirect characteristics (NUTS ^b level, size class, NACE ^c), but only rare units can be identified	Respondents cannot be identified
Access by	Only limited number of NSA ^a staff	NSA ^a staff	Researchers	Researchers	All
Data availability	Closed NSA ^a production environment	Closed NSA ^a production environment	Dedicated controlled NSA ^a environment for researchers	Data used outside NSA ^a environment, in the premises of the research entities	Data usually available on the NSA ^a website

^aNSA or Eurostat

^bNUTS level defines the geographical level of aggregation

^cNACE categories define the economic activities

the researcher may not need regional details but is interested in the exact age of individuals, whereas the standard files usually provide a medium level of regional details and age in bands).

The scientific-use files released by Eurostat are standard, i.e. they are prepared once for all access requests. Production of tailor-made files would be too burdensome, as the SDC protection measures must be always agreed with the NSAs.

Example of partial anonymisation methods for EU Labour Force Survey (LFS) scientific-use files:

AGE—by 5-year bands

NATIONALITY/COUNTRY OF BIRTH—up to 15 predefined groups

NACE (economic activity)—at 1-digit level

ISCO (occupation)—at 3-digit level

INCOME—provided only as (national) deciles and from 2009

HHNUM—household numbers are randomised per dataset, so that respondents cannot be tracked across time

The most common **SDC methods to anonymise (partially or completely) the microdata files** are:

- Recoding: provision of information at the more general level (e.g. age bands instead of exact age).
- Micro-aggregation: replacement of the original value of the variable (e.g. income) with the average of some (usually 3–5) similar units.
- Record swapping: swapping of, for example, persons between similar households. Swapping adds uncertainty about the identity of the unit in a microdata file.
- Rounding: replacement of original value with rounded figure.
- (Local) suppression: removal of identifying variables in the record or the entire record (e.g. a very large household).
- Sampling: provision of sampled microdata to increase uncertainty about identification as a record referring to particular individual may but does not have to be included in the sample.

The modes of access to secure-use files and scientific-use files are presented in Table 2.

The modes of access listed in Table 2 are complementary and some NSAs provide all options. As the operational costs may be high, the NSA services are sometimes payable.

Table 2 Modes of access to confidential data and respective protection measures

Microdata type used	Mode of access	Confidentiality protection
Secure-use files	<ul style="list-style-type: none">– On site (separate room, usually in the premises of the NSA where access is provided; researchers can see the data but have no internet access and cannot download or copy the data)	The final results of data analysis are checked for confidentiality (output checking). Each output is checked separately. In some systems, only a sample of output is checked; in others, researchers do it themselves
	<ul style="list-style-type: none">– Remote access (same functionalities as for on-site access but facilities provided online, no need to travel to the NSA)	
	<ul style="list-style-type: none">– Remote execution (authorised users submit codes that are executed on the data, but users do not see the data)	
	<ul style="list-style-type: none">– There might be also combinations of the above modes of access	
Scientific-use files	Files are sent to authorised researchers and used in the premises of research entities	The data are protected (partial anonymisation) before being sent to researchers. Researchers must ensure that the published results do not contain confidential data

5 Use Case: Access to European Statistical System Microdata (European Microdata)

How does the microdata access system work in practice? Eurostat applies a two-step procedure to grant access to microdata for research purposes. In the first step, organisations interested in accessing European microdata submit an application for recognition to Eurostat. In the second step, researchers from recognised research entities submit their concrete research proposals.⁴

Step 1 Recognition as a Research Entity

The recognition of research entities aims at identifying those organisations (or specific departments of the organisations) that carry out research and can be entrusted with confidential data. The assessment criteria refer to the purpose of the entity, its available list of publications and scientific independence. The entities must also describe security measures in place for microdata protection.

The content of the application is evaluated by Eurostat. Upon positive assessment, the head of a recognised research entity signs the commitment that the microdata will be used and protected according to the terms agreed. Eurostat publishes the list of recognised research entities on its website.⁵

To date (2017) more than 700 research entities were recognised. The majority of them are universities and research organisations (see Fig. 4).

Recognition of research entities was introduced by Eurostat to provide a contractual link with the legal entities, rather than with individual researchers.⁶

Step 2 Submission of Research Proposal

In the second step, researchers from recognised entities submit their concrete research proposals to Eurostat. Eurostat then consults all national statistical authorities that provided the data. If an NSA refuses the access, the data of that country are removed from the microdata file.

To be eligible, the research proposal must specify the scientific purpose of the research in sufficient detail, justify the need to use microdata and present the expected outcomes of the research. The results of the research must be made public. Each researcher named in the research proposal as a potential user of the microdata signs an individual confidentiality declaration, in which he or she commits to respect the specific terms of use of confidential data.

In the research proposal, researchers choose the microdata collections they are interested in. In 2017 Eurostat granted access to microdata to 12 data collections (see

⁴The legal basis for access to ESS microdata is Commission Regulation (EU) No. 557/2013 on access to confidential data for scientific purposes. The Regulation defines criteria for eligible research entities and research proposals. It also describes how the microdata shall be made available to researchers (modes of access).

⁵<http://ec.europa.eu/eurostat/documents/203647/771732/Recognised-research-entities.pdf>.

⁶However, in some national systems, only individual researchers are 'recognised'.

Fig. 4 Types of recognised research entities (in 2017)

Annex 1). Most of the European microdatasets are released as scientific-use files.⁷ The datasets most frequently demanded by researchers are EU Statistics on Income and Living Conditions (EU-SILC) and Labour Force Survey (LFS). Together they account for more than 70% of all access requests.

When the research proposal is accepted, the data are made available to the researchers. Researchers may access the data for the period specified in the research proposal. If so requested, researchers receive new releases of the approved microdatasets.

Once the project is finalised, researchers send Eurostat the resulting publications, which are made available on the dedicated website.⁸ Researchers must also destroy the confidential data received.

Eurostat receives around 350 applications for access to microdata per year.

6 Conclusions

The ESS microdata access system is specific as it creates a single entry point of access to European microdata owned by the NSAs. NSAs agree on the general access conditions (Regulation 557/2013) and are directly involved in decisions on the release of particular datasets in particular ways (anonymisation method and mode of access), and for particular projects (all NSAs are consulted about each access request).

⁷The anonymisation methods and/or output checking rules are agreed with NSAs.

⁸The publications issued using ESS microdata are available here: https://ec.europa.eu/eurostat/cros/content/publications-received_en.

For Eurostat, access to microdata has become a well-established process. Recently, Eurostat worked on modernising the microdata access system, e.g. launching online forms for microdata access applications and piloting online transmission of scientific-use files. The future plans aim to develop remote execution and to publish more public-use files.⁹ Closer collaboration with organisations such as CESSDA (Consortium of European Social Science Data Archives) should contribute to the improvement of microdata access services provided by Eurostat.

Annex 1: European Microdatasets Available for Scientific Purposes

European microdatasets available at Eurostat	Reference years, frequency of new releases	Business (B)/social (S) survey	Microdata file type	Mode of access
1. Adult Education Survey (AES)	2007, 2011	S	Scientific-use file	Off site
2. Community Innovation Survey (CIS)	2002–2012 (bi-annual)	B	Secure-use file and Scientific use file	On site (safe centre in Eurostat) and off site
3. Community Statistics on Information Society (CSIS)	2008–2014 (yearly)	S	Scientific-use file	Off site
4. Continuing Vocational Training Survey (CVTS)	2005, 2010	B	Scientific-use file	Off site
5. European Community Household Panel (ECHP)	1994–2001 (annual)	S	Scientific-use file	Off site
6. European Health Interview Survey (EHIS)	2006–2009 (one data collection depending on the country)	S	Scientific-use file	Off site

(continued)

⁹Currently available European public-use files are published here: https://ec.europa.eu/eurostat/cros/content/puf-public-use-files_en.

European microdatasets available at Eurostat	Reference years, frequency of new releases	Business (B)/social (S) survey	Microdata file type	Mode of access
7. European Road Freight Transport Survey (ERFT)	2011–2014 (annual)	B	Scientific-use file	Off site
8. European Union Statistics on Income and Living Conditions (EU-SILC)	2004–2015 (annual)	S	Scientific-use file	Off site
9. Household Budget Survey (HBS)	2010	S	Scientific-use file	Off site
10. Labour Force Survey (LFS)	1983–2015 (yearly)	S	Scientific-use file	Off site
11. Linked micro-aggregated data on ICT usage, innovation and economic performance in enterprises	2000–2010 ^a	B	Secure-use file	On site (safe centre in Eurostat)
12. Structure of Earnings Survey (SES)	1995, 2002, 2006, 2010, 2014	B and S	Secure-use file and scientific-use file	On site (safe centre in Eurostat) and off site

^aThe years covered by the MMD datasets vary from one country to another and are subject mainly to the availability of the Community Innovation Survey and Survey on ICT Usage and e-Commerce in Enterprises data

Aleksandra Bujnowska is a Statistical Officer in Unit B1 ‘Methodology and Corporate Architecture’ at Eurostat. She is leading a team ‘Statistical confidentiality and access to microdata’. For many years, she has been contributing to the development of the European microdata access system and has made several interventions on this subject at various events. She has also coordinated numerous European projects aiming at wider access to confidential data for scientific purposes and at efficient way of micro- and tabular data protection.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.