Glossary

Terminology on the provision of official statistical microdata for research purposes differs across countries, partly as a result of differences in national languages and cultures, partly as a result of the different structures of national statistical systems and of the different missions of relevant organisations.

The terminology adopted in the country factsheets is aligned with the wording used in the most recent European regulation of access to confidential data for scientific purposes (EC 557/2013 of 17 June 2013). Similar language choices have been made in a recent report of the OECD's Expert Group for International Collaboration on Micro-data Access (OECD 2014).

The table below groups relevant concepts thematically, offering for each of them: a definition: the source(s) from which it has been derived, and whether it is original or an adaptation; and any comments that are relevant to better understand the contents of the country factsheets.

The object: data concepts			
	Definition	**Sources**	**Comments**
Data	The physical representation of information in a manner suitable for communication, interpretation, or processing by human beings or by automatic means.	UNECE 2000, OECD 2014	
Database	A data file or set of data, typically machine-readable, with relationships expressed among data. Data stored in the database are independent of any particular application.	Adapted from UNECE 2000, OECD 2014	
Data set	Any organised collection of data.	UNECE 2000, OECD 2014	
Microdata	Data file where each record represents an individual statistical unit – a person, household, business or other entity.	Adapted from UNECE 2007, OECD 2014	
The context: the statistical system and statistical institutes			
	Definition	**Sources**	**Comments**
National Statistical Office (NSO) or National Statistical Institute (NSI)	The leading statistical agency within a national statistical system, in charge of collecting information and producing statistical analyses to support policy-making. In Europe, it is also the point of contact between Eurostat and the member state.	Adapted from OECD 2002a, OECD 2014	

The context: the statistical system and statistical institutes			
	Definition	**Sources**	**Comments**
Official statistics (or official data)	Data produced and disseminated by the national statistical system, excepting those that are explicitly stated not to be official. Official data may be collected directly by the NSO/NSI (typically, through census and surveys) or obtained from other sources, such as administrative records and registers (especially in the Nordic countries).	Adapted from OECD 2002a, OECD 2014	The definition varies slightly across countries. For example in the United Kingdom, official statistics is understood to be the statistics produced by the Office for National Statistics (ONS) plus the statistics produced by any other organisations involved in providing a public service; in France, where the word "public" is used instead of "official", the definition includes all data productions originating from statistical surveys of Institut National de la Statistique et des Études Économiques (INSEE), and the use of data collected by all organisations with a public service mission. The country factsheets include only information on data produced and disseminated by the NSI/NSO of each country, with the exception of Germany, for which we have also included the important Institut für Arbeitsmarkt- und Berufsforschung (IAB).
The users: research and researchers			
	Definition	**Sources**	**Comments**
Research	In the present context, this broadly refers to use of official data / microdata to explore, reveal and explain the social, economic, health, or environmental information within the data in original ways that bring out new knowledge.	DwB 2014	
Research purposes	European laws and regulations, as well as the statistical laws of individual member states, recognise the data needs of research and include provisions to offer (regulated) access to official microdata, including confidential data, for research purposes. Typically, these data would not be accessible to anyone outside the NSI that collected and processed them: in this sense, research purposes constitute an exception.	Adapted from DwB 2014	Most statistical legislations require research purposes to be non-commercial, some of them demanding evidence of that such as institutional mission of applicant's employer organisation or publication of outputs. Cognate activities that may or may not be recognised as research, with differences across countries, include public policy evaluation as well as teaching and learning. Please check the individual country factsheets for details.

The users: research and researchers			
	Definition	**Sources**	**Comments**
Researcher	Professional engaged in research to design or create new knowledge, products, processes, methods and systems. In the present context, the term refers primarily to a user of data for research purposes.	Adapted from UNECE 2007, UNECE 2009, OECD 2002b, OECD 2014	The definition varies across countries as some restrict it to members of recognised academic institutions such as universities (e.g. Germany, the Netherlands), while others extend it also to independent researchers and those working in government agencies, non-profits, international agencies (e.g. United Kingdom); the private sector is usually not included. Some countries include researchers working towards a PhD (e.g. Italy, Romania), post-graduate students (e.g. Estonia, France) and even undergraduates (e.g. Germany), though often under more restrictive conditions; others do not (e.g. Czech Republic). Most countries extend their definition to non-resident researchers, though additional restrictions (in terms of conditions of access, types of files that can be accessed, or evidence to be provided in support of the application) often apply. Please check the individual country factsheets for details.
Non-resident researcher	A researcher resident in, or affiliated to an institution of, a country different from the data provider's country.	Adapted from DwB 2014	The information provided in the factsheets is restricted to non-resident researchers from European countries (incl. EU member states and EFTA countries). This is because non-European countries have different data protection laws and therefore, different provisions often apply.
The risks: confidentiality and disclosure			
	Definition	**Sources**	**Comments**
Confidential (micro)data	Microdata which allow statistical units to be identified, either directly or indirectly, thereby disclosing individual information.	Adapted from EC 2009, EU 2013, OECD 2014	Microdata with direct identifiers such as names and addresses are hardly, if ever, released to researchers. Confidential data for scientific purposes are typically files from which direct identifiers have been removed, and which allow only for indirect identification of statistical units (for example, through some infrequent attribute or a combination of attributes). In practice, confidential data may be released in the form of either Secure-use files or Scientific-use files.

The risks: confidentiality and disclosure			
	Definition	**Sources**	**Comments**
Confidentiality	Obligation of the provider of information (data) to maintain the secrecy of that information.	UNECE 2009, OECD 2014	Confidentiality is a property of data that often results from legislative measures, typically a Statistical Law together with a Privacy Protection Law. While the latter is meant to protect the fundamental right of every individual to privacy, the former aims specifically to maintain the trust of respondents and ensure the long-term sustainability of official data collection efforts. The laws often attribute a specific status to national statisticians by submitting them to professional secrecy obligations. In this perspective, a tension appears between confidentiality protection and the data needs of researchers – a major reason why explicit provisions for research access are often included in legislation.
Disclosure	Disclosure relates to the inappropriate attribution of information to a data subject. It involves attribution (the association or disassociation of a particular attribute with a particular population unit) and identification of this statistical unit, directly or by indirect means (for example, a combination of attributes).	ESSNet SDC 2010, OECD 2014	
Disclosure risk	A situation in which an unacceptably narrow estimation of a respondent's confidential information is possible or if exact disclosure is possible with a high level of confidence.	ESSNet SDC 2010, OECD 2014	
Full / *de facto* anonymisation	Anonymisation is the set of methods applied to microdata in order to minimise the risk of identification of the statistical units concerned. With full anonymisation, re-identification of statistical units is almost impossible, while with *de facto* anonymisation, re-identification is not impossible but would take such a large investment of time, effort, or money that it is most unlikely to occur.	EC 2002, OECD 2014, DwB 2014	Anonymisation is a step forward relative to "de-identification", which defines the sheer removal of direct identifiers (OECD 2014). De-identified data may or may not be subsequently anonymised.

The risks: confidentiality and disclosure			
	Definition	**Sources**	**Comments**
Full / *de facto* anonymised (micro)data	Individual statistical records which have been modified in order to minimize, in accordance with current best practice, the risk of identification of the statistical units to which they relate.	EC 2002, OECD 2014	Fully anonymised data are not confidential and may thus be released as Public Use Files. *De facto* anonymised data are confidential and are often released as Scientific Use Files.
The solutions: modes of access			
	Definition	**Sources**	**Comments**
Public use files (PUFs)	These files contain microdata that: · allows users to make sensible inferences on the phenomenon for which the data were collected, and · has been subject to statistical methods and procedures that minimise disclosure risk and render the data non-confidential (fully anonymised).	Adapted from OECD 2014, DwB 2014	Due to their non-confidential (fully anonymised) nature and low level of disclosure risk, these data are often disseminated widely, not only to researchers but also to a broader public. They are either freely downloadable from the Internet (e.g. in Lithuania) or involve minimal restrictions such as registration and/or express agreement to terms of use (e.g. in Austria). The cost of production of these files is relatively high and their availability often limited (some countries not offering them at all, e.g. Poland), though growing. The countries that do produce them, do not always call them PUFs: for example the terminology is “Standardisierte Datensätze” in Austria, and “mIcro.STAT files” in Italy. The documentation is available in the national languages, but English translations are not always available. Please check the individual country factsheets for details. These files need to be distinguished from so-called Campus files (sometimes called Teaching files) which share the same form of provisions to users, but due to the heavy procedures used to make them non-confidential, they are not meant for proper statistical analysis of the phenomenon of interest, and can be used solely to teach statistical or data analysis methods (OECD 2014). Please check provider’s website (using links in country factsheets) for details.

The solutions: modes of access			
	Definition	**Sources**	**Comments**
Scientific use files (SUFs)	Confidential data for scientific purposes to which procedures have been applied to reduce to an appropriate level, and in accordance with current best practice, the risk of identification of the statistical unit.	Adapted from EU 2013, OECD 2014	These data files present an intermediate level of disclosure risk (*de facto* anonymised) and are often physically delivered to trusted researchers after signature of a contract or licence, for example on CD-Rom or USB device, or through an FTP server. Most countries offer SUFs, but the terminology varies: for example, they are called Research Data files - RDFs (Ireland), Micro-data Files for Research purposes - MFRs (Italy), or Fichiers de Production et Recherche - FPR (France). Please check the individual country factsheets for details. Notice that some countries also have an intermediate type of files: while their non-confidential nature allows them to be used by researchers together with a broader range of users (like PUFs), their modes and conditions of access as well as the application procedures, are similar to those of SUFs. This is the case for Anonymised Data Files - ADF (Ireland) and Standard Files (Italy).
Secure use files	Confidential data for scientific purposes from which direct identifiers have been removed, but which retain all other information. They thus display a high level of detail and a high disclosure risk.	Adapted from EU 2013, OECD 2014	These are typically released only to trusted researchers, typically upon signature of a formal contract and compulsory legal and ethical training. They can be accessed only under strictly controlled conditions, such as an onsite safe centre / data laboratory or a remote secure connection (remote access / remote execution), where the IT system disables download and copy of data, and allows output release only after appropriate checks for disclosure by NSI staff. Many European countries currently offer access to Secure-use files, but not all; and the offer of data provided through this mode is sometimes limited. Please check the individual country factsheets for details.

Modes of access to secure use files			
	Definition	**Sources**	**Comments**
Onsite safe centre	A data laboratory, usually on the premises of the data producer or provider, where access to Secure use files is offered under controlled conditions, both in terms of physical and IT security – for example, data cannot be downloaded or printed and output is subject to checks.	Adapted from ESSNet SDC 2010, OECD 2014, DwB 2014	Only a limited number of European countries currently have onsite safe centres (e.g. Italy, Hungary, the Netherlands), with a varying terminology: "Data enclave", "Onsite data facility", "Research Room" etc. Please check the individual country factsheets for details.
Secure remote execution / remote access	An IT solution that shares similar characteristics to onsite safe centres in terms of security settings (in particular, impossibility to download data), but it allows researchers to connect remotely to a central data server from their home institution. It relies on a secure connection between the server at the NSIs and a computer of the researcher, using firewalls and encryption techniques. Further procedures to control the login procedure like software tokens or biometrics are often added. One may further distinguish between remote execution (researchers do not actually see the data they are working with, and are only allowed to submit batch jobs that are executed by statistical agencies, and to receive outputs after a confidentiality check) and remote access (where researchers actually see, and work directly with, the data).	Adapted from ESSNet SDC 2010, OECD 2014, DwB 2014	Remote access / execution offer the flexibility for researchers to access a wide range of confidential data as in a data laboratory, while removing the constraints and costs of travelling to the NSIs premises; they are thus more popular than onsite solutions. Researchers tend to prefer remote access to remote execution because of the more direct connection with the data that they offer. Only a limited number of European NSIs currently offer remote solutions (e.g. Denmark, France, Germany, UK), though their number is growing fast. Please check the individual country factsheets for details.

References

DwB 2014: Data without Boundaries Work-Package 3, Deliverable D3.1, “Researcher accreditation - current practice, essential features, and a future standard”, Chapter 1 “Discovery”, pp. 6-46.

EC 2002: Regulation No 831/2002 on access to confidential data for scientific purposes, http://eur-lex.europa.eu/LexUriServ/site/en/oj/2002/l_133/l_13320020518en00070009.pdf

EC 2009: Regulation (EC) No 223/2009 of the European Parliament and of the Council on European statistics, http://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:087:0164:01:EN:HTML

EU 2013: Regulation (EU) No 557/2013 on access to confidential data for scientific purposes, http://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:164:0016:0019:EN:PDF

ESSNet SDC 2010: A network of Excellence in the European Statistical System in the field of Statistical Disclosure Control – ESSNet SDC 2010, Handbook on Statistical Disclosure Control, Version 1.2, http://neon.vb.cbs.nl/casc/.%5Csdc_handbook.pdf

OECD 2002a: Measuring the Non-Observed Economy: A Handbook, OECD, IMF, ILO, Interstate Statistical Committee of the Commonwealth of Independent States, 2002, Annex 2, Glossary, http://www.oecd.org/dataoecd/9/20/1963116.pdf

OECD 2002b Frascati Manual: Proposed Standard Practice for Surveys on Research and Experimental Development, 6th edition, OECD, 2002.

OECD 2014: Final report of the Expert Group for International Collaboration on Microdata Access, STD/CSSP/RD(2014)2.

UNECE 2000: Conference of European Statisticians Statistical Standards and Studies, No. 53, “Terminology on Statistical Metadata”, Geneva, 2000, http://www.unece.org/stats/publications/53metadaterminology.pdf

UNECE 2007: Managing Statistical Confidentiality & Microdata Access - Principles and Guidelines of Good Practice, 2007, p. 1, http://www.unece.org/fileadmin/DAM/stats/publications/Managing.statistical.confidentiality.and.microdata.access.pdf

UNECE 2009: Principles and Guidelines on Confidentiality Aspects of Data Integration Undertaken for Statistical or Related Research Purposes, Geneva, 2009, http://www.unece.org/fileadmin/DAM/stats/publications/Confidentiality aspects dataintegration.pdf