I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11092

Microdata for Social Sciences and Policy Evaluation as a Public Good

Ugo Trivellato

OCTOBER 2017

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11092

Microdata for Social Sciences and Policy Evaluation as a Public Good

Ugo Trivellato
FBK-IRVAPP, University of Padova, IZA and CESifo

OCTOBER 2017

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Microdata for Social Sciences and Policy Evaluation as a Public Good*

The balance between the right to privacy and the right to freedom of information is altered when scientific research comes into play, because of its inherent needs and societal function. This paper argues that, for research purposes, microdata should be characterised as a public good. The evolution of the rules and practices in the European Union (EU) for protecting confidentiality while allowing access to microdata for research purposes is reviewed. Two key directions are identified for further improvement: remote access to confidential data and the enlargement of the notion of 'European statistics' to include microdata produced for evaluating interventions (co)financed by the EU.

JEL Classification: C81, D04, H41, I38, J08, L5

Keywords: counterfactual impact evaluation, privacy and data protection, microdata access, intelligent openness, public policies

Corresponding author:
Ugo Trivellato
FBK-IRVAPP
Via S. Croce, 77
38122 Trento
Italy
E-mail: trivell@stat.unipd.it

* Forthcoming in: Crato N. and P. Paruolo (Eds), Data-driven Policy Impact Evaluation: How Microdata is Transforming Policy Design, Springer, 2018.

1. Setting the scene[*]

The issue of access to microdata for research purposes is multifaceted. In fact, it is at the crossroads of two concerns: the right to privacy, on the one hand, and the needs of scientific research on the other. The right to privacy is established in the European Convention on Human Rights and Fundamental Freedoms, reiterated and explicitly extended to the 'protection of personal data' in the Charter of Fundamental Rights of the European Union (EU).[1] However, this is not an absolute right, as it must be balanced against other competing rights: (i) freedom of expression and information, including freedom 'to receive and impart information' (Article 11), where freedom to receive information is considered to imply freedom to seek it; and (ii) freedom of the arts and sciences, which affirms that 'scientific research shall be free of constraint' (Article 13).

What are the data needs of scientific research per se, and for its role in improving the well-being of society? As the Royal Society (2012, p. 8) convincingly argues, 'open inquiry is at the heart of the scientific enterprise. [It] requires effective communication through … *intelligent openness*: data must be *accessible* and readily located; they must be *intelligible* to those who wish to scrutinise them; data must be *assessable* so that judgments can be made about their reliability …; and they must be *usable* by others. For data to meet these requirements it must be supported by explanatory metadata (data about data)' (emphasis added).

Economic and social sciences[2] face these concerns when data include information primarily on an identified or identifiable person, but also on another identified or identifiable agent, such as a firm or an administration.[3] How can these tensions be reconciled? This chapter will take the point of view of an EU-based researcher, focusing on some fundamentals of the issue and their policy implications, rather than on legal and technical aspects.

The rest of the chapter is organised as follows. Section 2 discusses the needs of scientific research and its societal role, in relation to processing microdata. Section 3 summarises the legislation on data protection. Section 4 reviews the evolution of the rules and practices for protecting confidentiality while allowing access to appropriate microdata for research purposes. Section 5 discusses the present state of play in the EU as a whole. The concluding section focuses on the way forward.

2. Scientific research: intrinsic needs and societal role

[*] The author wishes to thank participants at the seminar held at CRIE-JRC, Ispra, 18 November 2016, for a stimulating discussion, and the editors for comments and suggestions on an earlier draft.

[1] See Council of Europe (1950, Article 8) and European Parliament (2000, Articles 7 and 8), respectively. The Convention was ratified by all member states of the Council of Europe, among which are those of the EU; the Charter became legally binding with the entry into force of the Treaty of Lisbon, 1 December 2009.

[2] Biological and medical sciences frequently involve personalised intervention, and thus face additional challenges, which are out of the scope of this chapter.

[3] In most countries, and in the EU, a concern for confidentiality also extends to the information provided by other units: enterprises, administrations, other institutions, associations, etc. (European Statistical System Committee 2011).

This section outlines the role of individual information in scientific research, points to the growing need for microdata for social science and policy evaluation, and stresses the importance of replicability in science. These points are discussed in turn, and lead to a characterisation of microdata as a public good, i.e. a non-excludable and non-rivalrous good.

First, the distinctive feature of scientific research is the collective use of individual data. This is elucidated in Recommendation No R (97) of the Council of Europe on the protection of personal data collected and processed for statistical purposes (Council of Europe 1997a).[4] It considers statistics as a scientific discipline that, starting with the basic material in the form of individual information about many different persons, elaborates 'statistical results', understood as characterising 'a collective phenomenon'. This interpretation is extended to fundamental scientific research, which 'uses statistics as one of a variety of means of promoting the advance of knowledge. Indeed, scientific knowledge consists in establishing permanent principles, laws of behaviour or patterns of causality [or patterns of a phenomenon] which transcend all the individuals to whom they apply' (Council of Europe 1997b, p. 7). Moreover, the recommendation points to the need in both the public and private sectors for reliable statistics and scientific research (i) for analysing and understanding contemporary society and (ii) for evidence-based decisions. Summing up, statistics and scientific research separate the information from the person: personal data are processed with a view to producing consolidated and anonymous results.

Second, scientific research is experiencing an increasing trend in the use of microdata. Various factors operate to bring about this trend. Some of them act from the supply side, such as technological and statistical advances in data processing, which are making databases of individuals, households and firms more and more widely available. On the demand side, two factors are largely contributing to this trend: (i) from an analytical perspective, the increasing attention paid to individuals (broadly agents), their heterogeneity, micro-dynamics and interdependencies; and (ii) the focus on distributive features of policies and on specific target groups of agents, such as in welfare policies and active labour market policies.

In this area, there is a strong demand for assessing the causal effects of programmes, i.e. for estimating the effect of policies on outcomes: this is the core aim of counterfactual impact evaluation (CIE).[5] Correct causal inference depends on knowledge of the characteristics of the population members – the treated and the control groups – relevant to the selection process, and the availability of adequate data on them.

The third aspect has to do with replicability, which is essential to science: researchers should be able to re-work analyses and challenge previous results using the same data (Royal Society 2012, pp. 26–29). Along the same line, and also dealing with CIE, Heckman and Smith (1995, p. 93)

[4] A recommendation is a legal instrument of the Council of Europe, as well as forum organisations such as the OECD and UNECE, that is not legally binding but through the long-standing practice of the member countries is considered to have great moral force. This type of instrument is often referred to as 'soft law'.

[5] The literature on CIE is huge. See, recently, Athey and Imbens (2017). See also European Commission (2013b), a guide for officials responsible for the implementation of European Social Fund-funded interventions.

stress that 'evaluations build on cumulative knowledge'. Science is an incremental process that relies on open discussion and on competition between alternative explanations. This holds both for fundamental research and for policy research, and implies access to microdata – possibly personal data.

Can the peculiar 'good' of personal data processed for research purposes therefore be characterised as a public one?

To answer the question, first consider official statistics, compiled and disseminated by official statistical agencies. Official statistics are a non-rivalrous good. Moreover, collective fixed costs have a dominant role in producing them (Malinvaud 1987, pp. 197–198). But it is not a public good per se, as it is excludable: it would be possible to discriminate among users, both through pricing and through selective access. Thus, characterising official statistics as a public good is a normative issue, the result of a choice in a democratic society. Currently this is a common view: official statistics need to be (and in many countries are) a public good.

Among other things, this view is supported by the principle of 'impartiality', one of the Fundamental Principles of Official Statistics adopted by the United Nations Statistical Commission for Europe (UNECE) (UNECE 1992),[6] as well as of the principles for European statistics set out in European Parliament (2009). As stated in the latter, ' "impartiality" [means] that statistics must be developed, produced and disseminated in a neutral manner, and that all users must be given equal treatment'.

This argument can be extended to microdata for research purposes, especially when microdata come from public sources or funding (Wagner 1999, among others),[7] provided that: (i) eligibility of access is restricted to research purposes, and in appropriate ways to researchers; and (ii) data access does not compromise the level of protection that personal data require. While intelligent openness remains the paradigm (Royal Society 2012, p. 12), the operational solutions required to achieve it safely remain an issue.

3. EU legislation on data protection

The starting point is Directive 95/46/EC 'on the protection of individuals with regard to the processing of personal data and on the free movement of such data' (European Parliament 1995; Directive hereafter).[8] Among its features, two are worth considering.

- Like all EU directives, Directive 95/46/EC is addressed to the Member States and requires them to achieve a result – data protection – without dictating the exact means for fulfilling it, thus

[6] The Fundamental Principles of Official Statistics, initially adopted by UNECE 'in the region of the Economic Commission for Europe', were adopted by the United Nations (UN) Statistical Commission in 1994 and endorsed by the UN General Assembly in 2014.

[7] The contrast between official statistics and microdata has lessened substantially over the past decade.

[8] From 25 May 2018 the Directive will be replaced by Regulation (EU) 2016/679 (European Parliament 2016). The new Regulation does not appreciably change the rules for processing personal data for scientific or statistical purposes. Its essential innovation is that, like all EU regulations, it is immediately enforceable as law in all Member States and is overseen by the judicial authority of the European Court of Justice.

leaving some leeway. It is up to the Member States to bring into force the national law(s) and the administrative provision(s) necessary to comply with the Directive.

- With regard to its scope, the Directive deals with data protection at large, covering almost all kinds of personal data and all of their uses. Thus, it is sparing in offering provisions for their processing for 'statistical or research purposes'.[9]

The Directive states that '"personal data" shall mean any information relating to an identified or identifiable natural person', where an identifiable person is one who can be identified directly (e.g. by reference to an identification number) or indirectly (i.e. by reference to data on one or more factors specific to his or her physical, physiological, economic, cultural or social identity).

As for the general provisions of the Directive, they stipulate that personal data must be: (i) processed fairly and lawfully; (ii) collected for specified, explicit and legitimate purposes, ordinarily with the informed and unambiguous consent of the person concerned; (iii) adequate, relevant and not excessive in relation to the purposes for which they are collected and/or further processed; (iv) accurate and, where necessary, kept up to date; and (v) kept in a form that permits identification of data subjects for no longer than is necessary for the purposes for which the data were collected or further processed. In addition, the Directive establishes: (vi) the information to be given to a person, where data have not been obtained from him or her; (vii) some rights of the person, chiefly the right to access to his or her personal data and rectify them; and (viii) the need for technical and organisational measures to be taken to ensure the secure processing of personal data.

When personal data are processed for statistical or research purposes, the Directive determines some derogations to the general provisions, provided that Member States put in place appropriate safeguards. The crucial exemption states that it is possible to further process for statistical or research purposes data that were collected for other specific and legitimate purposes. Additional waivers apply to points (v)–(vii) above.

Clearly, substantial room is left to the Member States when transposing the Directive into national legislation, and to EU institutions for European legislation.

Member States differ appreciably with respect to infrastructure for data collection and dissemination (e.g. national statistical institutes (NSIs) and other statistical authorities and/or social science data archives (DAs), data sources – statistical surveys and/or administrative records –). Besides, countries differ with respect to the focus and intensity of the concerns for confidentiality and the ways of handling them, as they are rooted in each country's culture, legislation and practices (Trivellato 2000, pp. 676-681).

4. A cursory review of data access for research purposes in the EU

At the EU level, the process was quite laborious and took a long time, over two rounds: from 1997 to 2002, and from 2009 to 2013. In each round, two regulations were adopted. Note that, in contrast

[9] The Directive addresses 'historical, statistical or research purposes'. 'Historical purposes' are dropped as irrelevant in the present context.

with directives, regulations are binding and directly applicable in all Member States.

Council Regulation No 322/97 (Council of the EU 1997) established the initial framework for the production and dissemination of European statistics,[10] as well as for microdata access for research purposes. On the latter, it states:

(a) 'To determine whether a statistical unit is identifiable, account shall be taken of all the means that might *reasonably* be used by a third party to identify the statistical unit'.[11] This does not imply a zero risk of identification; rather, the risk is considered to be practically non-existent when identification would require overly complicated, lengthy or costly operations. Obviously, when statistical units are not identifiable the microdata set is considered anonymised.

(b) Access to confidential data[12] transmitted by the national authorities to Eurostat may be granted by Eurostat itself, provided that it is for scientific purposes and under two further conditions: (b1) explicit approval from the national authority which supplied the data, and (b2) enactment of appropriate safeguards for the physical and logical protection of the data.

To draft the subsequent regulation specifically on access to confidential data for scientific purposes, Eurostat was active in promoting an informed debate, with the involvement of NSIs and of members of the research community in advisory committees and working parties, and at conferences and seminars (e.g. Jenkins 1999, Wagner1999, Trivellato 2000, CEIES 2003).[13] Their contributions converged on a guiding principle: capitalising on technological developments and taking appropriate regulatory, organisational and administrative measures (including sanctions), microdata – possibly confidential microdata – should be made available to researchers in accordance with a principle of proportionality (i.e. they should be adequate and not excessive in relation to the purpose) and in a variety of formats. Formats range from 'safe data', i.e. anonymised microdata distributed as public-use files (PUFs), to confidential microdata just net of the identifier made accessible to researchers via a 'virtual safe setting',[14] i.e. via safe, remote online access to a secure data storage and computing laboratory within Eurostat, and/or a European data archive facility, under appropriate undertakings. In short, this guideline points to the implementation of an adequate set of safe open environments for analysing microdata for scientific purposes, at no (or marginal) cost, and with no appreciable risk of infringing confidentiality.[15]

[10] In accordance with the Maastricht Treaty in force up to 2009, the Regulation refers to 'the Community authority' and to 'Community statistics'. In this chapter the current wording is used: 'the Commission' or 'the Commission (Eurostat)' or 'Eurostat', where appropriate, and 'European statistics'.

[11] Moreover, the Regulation points out that data taken from sources which are available to the public are not considered confidential (Article 13; emphasis added).

[12] The term 'confidential data' was synonymous with 'personal data' until the promulgation of Regulation (EC) 831/2002, where it takes a quite restrictive meaning. Its meaning is further modified in Regulation No 223/2009.

[13] The European Advisory Committee on Statistical Information in the Economic and Social Spheres, better known under its French acronym CEIES, was set up in 1991 to assist the Council and the Commission in the coordination of the objectives of the EU's statistical information policy. It was replaced by the European Statistical Advisory Committee in 2008.

[14] This mode of access is also known as 'remote data access' or 'microdata online access'.

[15] Based on the experience at the United Kingdom Data Archive, for anonymised microdata Jenkins (1999, p. 78–81) advocates 'universal access […] for all *bona fide* non-commercial users, subject to registration and standard

CEIES (2002) gave significant support to this process: '1. Much significant research in the social and economic spheres, both fundamental and of relevance to the formulation and evaluation of public policies, can only be undertaken with microdata; it cannot be done using published statistics or aggregate records. … 9. CEIES recommends that Eurostat should establish the feasibility of a virtual safe setting as an alternative to a physical safe setting. If the virtual setting can be put into place, it will be much more cost effective and provide a preferred means of access for the research community'.

Eventually the Commission adopted Regulation (EC) No 831/2002 (Commission of the European Communities 2002), but it took a more conservative stance. Its stated aim was 'to establish, for the purpose of enabling statistical conclusions to be drawn for scientific purposes, the conditions under which access to confidential data transmitted to Eurostat may be granted'. It modified two crucial definitions of the 'father' Council Regulation No 322/97.

(a) It established that anonymised microdata shall mean 'individual statistical records which have been modified in order to *minimise*, in accordance with current best practice, the risk of identification of the statistical units' (Article 2, emphasis added). This is at odds with the Council Regulation's criterion based on 'all the means that might *reasonably* be used by a third party'.
(b) Previously microdata were considered confidential when they allowed statistical units to be identified, either directly or indirectly, while in this Regulation ' "confidential data" shall mean data which allow only indirect identification of the statistical units concerned' – which is sensible – and ' "access to confidential data" shall mean either access [to proper confidential data] on the premises of Eurostat *or release of anonymised microdata*' distributed under license (Article 2, emphasis added) – which is an inconsistent, restrictive adhockery.

The step back with respect to Council Regulation No 322/97 is apparent. Data access was restricted within the conservative paradigm that a balance has to be struck between two conflicting aims, privacy and data needs for scientific research .

The design of the two procedures envisaged in (b) had clear drawbacks. 'Safe data' had to pay the price of a substantial reduction in the information content of the datasets, with the addition burden of obtaining a license. The 'safe centre' on the premises of Eurostat paid the price of severe restrictions placed on access opportunities for researchers, because of the substantial direct and indirect costs incurred by them.[16]

Nonetheless, the availability of microdata from some surveys, granted under Regulation (EC) No 831/2002 via access to the safe centre, turned out to be a significant opportunity. It opened up

undertakings of non-abuse, at no cost'. Moreover, he points out that additional components are essential for a sound use of microdata for research purposes: extensive documentation and metadata; information, assistance and training; significant involvement and feedback from analytical users (e.g. via user groups and scientific boards of advisors).

[16] This feature demonstrably jeopardised the equality of opportunity of access for researchers. Substantially higher costs were incurred by researchers who travelled a considerable distance to reach the safe setting. Moreover, many academics who needed to combine research with teaching and other obligations were unable to stay at the safe centre long enough to conduct their research.

research to cross-country and (almost) Europe-wide comparisons on significant topics; it helped to create two-way trust between Eurostat and the community of analytical users; and it contributed to stimulating a growing demand for microdata in the economic and social domains, and to pushing forward demand for integration of data from different sources and along the time dimension (e.g. employer–employee linked longitudinal data).

Moreover, various initiatives by Eurostat, the Organisation for Economic Co-operation and Development (OECD) and UNECE have offered new insights on data access. While advances in computer processing capacity, in record linkage and statistical marching open new opportunities for indirect identification, similar developments are also taking place for secure online data access, statistical disclosure control, database protection, disclosure vetting procedures, etc. Overall advances in information and communications technology (ICT) can be harnessed to provide a secure, monitored and controlled environment for access to microdata (e.g. UNECE 2007).

Meanwhile, new potential was emerging from the use of administrative records. Registers draw on the entire (administratively defined) population, are regularly updated and come at no direct cost, which allows the enhancement of statistical and research results (Eurostat 1997, European Commission 2003; see also the series of European Statistical System (ESS) ESSnet projects at https://ec.europa.eu/eurostat/cros/page/essnet_en).

The focus on research data from public funding was another stimulating perspective. In January 2004, the ministers of science and technology of the OECD member countries adopted a *Declaration of access to research data from public funding* and invited the OECD to develop 'a set of guidelines based on commonly agreed principles to facilitate cost-effective access to digital research data from public funding'. The principles and guidelines, drafted by a group of experts after an extensive consultation process, were endorsed by the OECD Council and attached to an OECD recommendation (OECD 2007).

Lastly, a notable impetus to revision of the EU legislation and practices came from advances made in some Member States, such as the Netherlands, the Nordic countries and the United Kingdom (UK). In addition to the diversified practices of safe data dissemination and the establishment of safe centres, between 2000 and 2010 there was a move from piloting to implementation of remote data access services. They include:

(a) 'Remote execution'. Registered researchers submit their command files to the safe centre via email, and written in one of the admissible statistical packages. At the centre they are moved across a firewall to the server holding the data, and the tasks are run. The results, in the form of output from analyses, are then returned to the researcher by email. This is the case at the LIS Data Center in Luxembourg (home of the Luxembourg Income Study (LIS) and the Luxembourg Wealth Study (LWS) databases; see http://www.lisdatacenter.org) and at some other organisations, including IAB-FDZ, the Research Data Center of the German Employment Agency at the Institute for Employment Research. Note that this mode of access may be cost effective, but it severely limits the level of interaction of the analyst with the data.

(b) 'Decentralised' access to a safe centre. Under this mode, accredited researchers, in addition to accessing the data at the safe centre, can do so from 'safe rooms' in (a moderate number of) offices which are part of the data provider's network or at selected universities and other research institutions. For instance, this is the case for the initial provision of decentralised data access in Denmark (Andersen 2003).

(c) Proper 'remote data access'.[17] While its basic format is common to the NSIs and DAs that launched it, procedures and practices vary appreciably in several respects: accreditation procedures, domain of the data made accessible, researcher authentication procedures, output checking, output release, etc.. Pivotal cases are remote access at Statistics Denmark (Statistics Denmark 2014, pp. 75–79) and the Microdata ON-line Access implemented at Statistics Sweden starting from the end of 2005 (Hjelm 2006).

Within this renewed interest in extending the accessibility of microdata, the European Parliament (2009) adopted a new Regulation on European statistics, No 223/2009. Known as the 'Statistical Law', it marks a profound change. First, it takes a broad, systematic approach to European statistics. It encompasses (i) the reformulation of statistical principles; (ii) the reshaping of statistical governance, centered around the notion of the ESS and the role of the ESS Committee in providing 'professional guidance to the ESS for developing, producing and disseminating European statistics'; (iii) the production of European statistics, with provision of access to administrative data sources; (iv) the dissemination of statistics; (v) and, finally, statistical confidentiality.

No less important are the novelties regarding data access. First, dealing with 'data on individual statistical units [that] may be disseminated in the form of a *public use file*', the new Regulation confirms the criterion of taking into account 'all relevant means that might *reasonably* be used by a third party' (Article 19; emphasis added). The very same notion of a PUF, and its placement under the heading of 'dissemination of European statistics', makes it clear that the set of anonymised data is complementary to the set of confidential data, and that provisions on data protection do not apply to the former.

Second, the boundary of the confidential data that researchers may access for scientific purposes is sensibly and neatly stated: 'Access may be granted by the Commission (Eurostat) to *confidential data which only allow for indirect identification of the statistical units*'.[18] On the other hand, it remains for the Commission to establish 'the modalities, rules and conditions for access at Community level' (Article 23, emphasis added).

Eurostat implemented various actions for the task and launched several ESSnet projects, such as the feasibility study 'Decentralised Access to EU Microdata Sets' and the project 'Decentralised

[17] As the number of offices with safe rooms increases, and safe rooms are extensively installed in universities and other research institutions, the distinction between decentralised access to a safe centre and remote data access patently fades out.

[18] If the data have been transmitted to Eurostat, the usual approval of the NSI or other national authority which provided them is required. Note that the restriction to grant access only to confidential data also applies to NSIs and other national authorities.

and Remote Access to Confidential Data in the ESS (DARA)', whose aim was to establish a secure channel from a safe centre within an NSI to the safe server at Eurostat, so that researchers could use EU confidential microdata in their own Member States.[19]

Two other important initiatives on transnational access to official microdata were the 'Data without Boundaries (DwB)' project, promoted by the Consortium of European Social Science Data Archives (CESSDA) and launched in May 2011 (http://www.dwbproject.org/), and the 'Expert Group for International Collaboration on Microdata Access', formed in 2011 by the OECD Committee for Statistics and Statistical Policy. The composition of the two teams – both with diversified competences, but also with some moderate overlapping – and the constant collaboration between DwB and Eurostat favoured cooperation. Significant results are in OECD (2014), Data without Boundaries (2015) and Jackson (2018).

Finally, Commission Regulation No 557/2013 was adopted (European Commission 2013a). This was a long way towards reach an appropriate legal framework for access to confidential data for research purposes .

5. The state of affairs in the EU

By combining the provisions of the two extant Regulations,[20] microdata files made available to researchers fit into three categories and four modes of access. The categories are:

(a) '*Public-use files*': sets of anonymised records of individual statistical units. Provisions for confidential data do not apply to these. On the other hand, no indications are given on how to disseminate them.

(b) '*Scientific-use files*': confidential 'data to which methods of statistical disclosure control have been applied to reduce to an appropriate level and in accordance with current best practice the risk of identification of the statistical unit'. Access is granted to researchers from Member States, European Economic Area (EEA)/European Free Trade Association (EFTA) countries and some EU candidate countries. It takes place in two steps: (b1) recognition of the institution as a research entity; and (b2) approval of a research project, submitted by researchers linked to the research entity, who also need to sign a confidentiality undertaking. 'Scientific-use files' are then transmitted to the research entity.

(c) '*Secure-use files*': confidential 'data to which no further methods of statistical disclosure control have been applied'. Patently these are the most informative, and arguably in many cases are of peculiar interest for research purposes. The accreditation procedure does not vary. But 'access to secure-use files may be granted provided that the results of the research are not released without prior checking to ensure that they do not reveal confidential data'.

Moreover, 'access to secure-use files may be provided only within Commission (Eurostat) access facilities or other access facilities accredited by the Commission (Eurostat) to provide access

[19] See https://ec.europa.eu/eurostat/cros/content/decentralised-and-remote-access-confidential-data-ess-dara_en and Gürke et al. (2012), Statistics Denmark (2014), Tubaro et al. (2015).

[20] Additional information is taken from European Commission (2016).

to secure-use files'. Considering that ' "access facilities" means the *physical or virtual* environment … where access to confidential data is provided', this implies that for secure-use files two modes of access are envisaged: (c1) at Eurostat's safe centre (or another accredited access facility); or (c2) via remote data access.

[Figure 1 about here]

Figure 1, adapted from OECD (2014, p. 8), sketches the secure open environments for accessing the microdata, where the four zones designate datasets with various levels of risk of identification, and with which different procedures are associated. Zone 0, the white area outside the circle, refers to anonymised datasets (PUFs), which present a negligible risk of re-identification and are made publicly available, subject to registration and possibly standard undertakings. Zone 1 designates the set of scientific-use files, which entail a moderate risk of identification; they are transmitted to recognised research entities, where they can be accessed by the accredited researchers under adequate security safeguards. Zone 2 designates the set of secure-use files: they entail a high risk of identification, and can be accessed only at the access facility itself or via remote data access. NSIs and other relevant national authorities provide directly identifiable personal data to Eurostat in Zone 3: access to them is restricted to Eurostat and the access facility, which perform the set of operations needed to make the confidential data available for research purposes.

This description refers to the 'law on the books'. What about its implementation? The essential results and plans are in Bujnowska (2015, 2016) and at the CROS (Collaboration in Research and Methodology for Official Statistics) Portal Group 'Microdata Access' at https://ec.europa.eu/eurostat/cros/content/microdata-access_en.

Focusing on confidential data,[21] priority has been given to the production and distribution of scientific-use files, also because they demand an extensive, dataset-specific application of statistical disclosure control methods. Results are quite satisfactory. A total of 11 microdata sets had been made available as of December 2016; some 580 research entities have been recognised, and since 2014 more than 300 research proposals per year have been submitted.

As for secure-use files, on-site access has been provided by Eurostat's safe centre in Luxembourg, active for decade, with a comparatively modest investment. Secure-use files are available for the 'Community Innovation Survey' and the 'Structure of Earnings Survey' (two of the 11 surveys for which scientific-use file versions have been provided), and for the 'Micro-Moments Dataset', an innovative linked micro-aggregated dataset on ICT usage, innovation and

[21] As for PUFs, the 'Public use files for Eurostat microdata' project was launched in 2005. In January 2017 Eurostat and CROS released PUFs for the EU Labour Force Survey (2012 and 2013) and the EU Statistics on Income and Living Conditions (2013), for Finland, Germany, Hungary, the Netherlands, and Slovenia. Access is open to those registered users of the CROS portal that have also applied for membership of the group on PUFs. Unfortunately, the files are prepared in such a way that individual entities cannot be identified (note that this practice is not in accordance with the identification criterion established by Regulation No 223/2009). This causes a loss in information value, and makes the PUFs useful essentially for training and testing only. Hopefully, this restrictive choice is in part – and might be further – compensated by provision of secure access to confidential data.

economic performance in enterprises, which enables studies of the economic impact of ICT at company level to be compared across a large sample of European countries.

The use of remote data access secure-use files is attractive for both researchers and Eurostat (or other accredited access facilities), since the microdata do not leave the facility and all output can be controlled. However, no significant advances have so far been made on that front. One crucial reason has been that the envisioned partnership between the ESS and CESSDA had to face a long delay, because of the prerequisite for CESSDA to be recognised as a European Research Infrastructure Consortium (ERIC).

6. Two suggestions for improvements

Current initiatives and further steps planned by Eurostat and the ESS for enhancing the use of microdata deal persuasively with various aspects. This section will focus on the need for improvements in two directions: remote data access to secure-use files, and reception of a suitably extended meaning of 'European statistics'.

It is no longer controversial that remote data access is an essential ingredient for providing a level playing field for scientific research and for supporting the EU's objective of a 'European research area in which researchers, scientific knowledge and technology circulate freely'[22]. Given the experience of NSIs and DAs in several countries, recently extended to transnational remote data access,[23] it is also largely accepted that remote data access is the most effective mode for sharing highly informative confidential data safely.

The good news is that in June 2017 CESSA became an ERIC. The opportunity for a partnership between Eurostat and CESSDA is now open. This should be a priority for the European Commission and Eurostat. Collaboration should be focused on the facility that will provide the entry point for the EU's microdata access system (possibly involving NSIs). It should also extend to essential additional components, such as: information on ESS microdata products – scientific-use and secure-use files, and any PUFs – (e.g. making them discoverable through the Resource Discovery Portal managed by CESSDA); metadata products and services; training and assistance; user conferences, and current involvement and feedback from researchers; and production of new microdata files especially for scientific research, which entails the integration of data from different sources – or archives – and along the time dimension.

The second area where there is a strong demand for improvement falls under the heading 'reception of a suitably extended meaning of European statistics'. First, Regulation No 223/2009 is clear on the need for a more intensive use of administrative records: Eurostat and NSIs 'shall have access to administrative data sources, from within their respective public administrative system, to the extent that these data are necessary for the development, production and dissemination of

[22] Article 179(1) of the Treaty on the Functioning of the EU.

[23] In addition to some pilots carried out within the DARA and DwB projects, it is worth mentioning the Nordic Microdata Access Network, which includes Denmark, Finland, Norway, Sweden, Greenland and Iceland (Statistics Denmark 2014, Thaulow and Nielsen 2015).

European statistics' (Article 24). This change may have started, and the production of statistics may also have moved to increased use of administrative sources. But such change is not reflected in increased access to this new data. As pointed out in OECD (2014, Executive summary, Recommendation 51), 'it [is] important to move the information base for microdata access files at the same pace as for statistical production when an office increases its use of administrative data'.

Second, microdata are also produced for monitoring and evaluation of interventions (co)financed by the EU. Their relevance is apparent, as evaluations (at large) are obligatory for all the European Structural and Investment Funds (European Commission 2015) and more emphasis has been placed on CIE, particularly for European Social Fund-funded interventions and research projects. Microdata resulting from interventions as well as from CIE research projects (co)financed by the EU will be made accessible as confidential data for research purposes, preferably as secure-use files via online access to an access facility.

This aim is motivated, and could be implemented, as follows.

1. Microdata produced for monitoring and evaluation should be recognised as part of 'European statistics', and hence included in the European statistical programme. In fact, European statistics are defined as 'relevant statistics necessary for the performance of the activities of the Community' and are 'determined in the European statistical programme' (Regulation No 223/2009, Article 1). Currently microdata produced for monitoring and evaluation are not included in the programme. By it is hardly reasonable to deny that they are 'relevant statistics necessary for the performance of the activities of the Community'.[24]
2. Organisations or research units receiving (co)-financing from the EU to carry out evaluations should supply to the European Commission, along with the final report, the full primary data produced, in an appropriate form (i.e. intelligible and assessable, with the relevant metadata). In accordance with the content and the planned use of the microdata, it will be up to the Commission to decide which unit they should deliver this to (e.g. a relevant Directorate-General, Eurostat or the Joint Research Centre).
3. The unit in charge of the management of the microdata should prepare the confidential files – preferably secure-use files – in accordance with the standards set by Eurostat.

It is not a trivial task to specify and implement the above proposal. In the author's opinion, it would be sensible to consider and discuss these steps promptly.

[24] Absurdly, how the Community would be justified to spend money for the production of microdata not relevant and necessary for its activities? From a substantive point of view, one should consider also two further reasons, already referred to, for making these research microdata publicly available: they are data from public funding (OECD 2007); the general argument of Royal Society (2012, p. 8) holds.

References

Athey S, Imbens GW (2017) The state of applied econometrics: causality and policy evaluation. J Econ Perspect 31(2):3–32.

Andersen O (2003) Access to micro data from Statistics Denmark. In CEIES (2003), cit., pp 147–152.

Bujnowska A (2015) Access to EU microdata for research purposes. Paper presented at the joint UNECE/Eurostat work session on statistical data confidentiality, Helsinki, Finland, 5–7 October 2015. http://www1.unece.org/stat/platform/display/SDCWS15/Statistical+Data+Confidentiality+Work+Session+Oct+2015+Home

Bujnowska A (2016) ESS microdata access – recent developments. European Commission, Brussels. https://circabc.europa.eu/sd/a/6663936c-6704-4fdb-97f1-bccb7608d297/Item%202.1%20Microdata%20access(0).pdf

CEIES (2002) Opinion given on Dissemination Policy: 'Access to microdata for research purposes'. Eurostat, Luxembourg.

CEIES (2003) 19th CEIES seminar: innovative solutions in providing access to microdata, Lisbon, 26–27 September 2002. Publications Office of the European Union, Luxembourg.

Commission of the European Communities (2002) Commission Regulation (EC) No 831/2002 of 16 May 2012 implementing Council Regulation (EC) No 322/97 on Community Statistics, concerning access to confidential data for scientific purposes. OJ L 133, 18.7.2002, pp 7–9.

Council of Europe (1950) European Convention for the Protection of Human Rights and Fundamental Freedoms. European Court of Human Rights, Council of Europe, Strasbourg. http://www.echr.coe.int/Documents/Convention_ENG.pdf

Council of Europe (1997a) Recommendation No R (97) 18 of the Committee of Ministers. Council of Europe, Strasbourg. https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680508d7e

Council of Europe (1997b) Explanatory Memorandum to Recommendation No R (97) 18 of the Committee of Ministers. Council of Europe, Strasbourg. https://rm.coe.int/CoERMPublicCommon SearchServices/DisplayDCTMContent?documentId=090000168050a58f

Council of the European Union (1997) Council Regulation (EC) No 322/97 of 17 February 1997 on Community Statistics. OJ L 52, 22.02.1997, pp 1–7.

Data without Boundaries (2015) EU Seventh Framework Programme, Project No 262608. http://www.dwbproject.org/

European Commission (2003) Business register recommendations manual: theme 4 – industry, trade and services. Publications Office of the European Union, Luxembourg.

European Commission (2013a) Commission Regulation (EU) No 557/2013. OJ L 1364, 18.6.2013, pp 16–23.

European Commission (2013b) Design and commissioning of counterfactual impact evaluations. Publications Office of the European Union, Luxembourg.

European Commission (2015) European structural and investment funds 2014–2020: official texts and commentaries. Publications Office of the European Union, Luxembourg.

European Commission (2016) Guidelines for the assessment of research entities, research proposals and access facilities, version 1.4. Publications Office of the European Union, Luxembourg. http://ec.europa.eu/eurostat/documents/203647/771732/guidelines-assessment.pdf.

European Parliament (1995) Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995. OJ L 281, 23.11.1995, pp 31–50.

European Parliament (2000) Charter of Fundamental Rights of the EU. OJ C 364, 18.12.2000, pp 1–22.

European Parliament (2009) Regulation (EC) No 223/2009 of the European Parliament and of the Council of 11 March 2009. OJ L 87, 31.3.2009, pp 164–173.

European Parliament (2016) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016. OJ L 119, 4.5.2016, pp 1–88.

European Statistical System Committee (2011). European statistics code of practice for the national and Community statistical authorities. Eurostat and European Statistical System, Luxembourg.

Eurostat (1997) Proceedings of the seminar on the use of administrative sources for statistical purposes, Luxembourg, 15–16 January 1997. Publications Office of the European Communities, Luxembourg.

Gürke C, Schiller D, Gadouche K (2012) Report on the state of the art of current safe centres in Europe. EU Seventh Framework Programme, Project No 262608, Data without Boundaries, Deliverable 4.1. http://www.dwbproject.org/export/sites/default/about/public_deliveraples/d4_1_current_sc_in_europe_report_full.pdf.

Heckman JJ, Smith JA (1995) Assessing the case for social experiments. J Econ Perspect 9(2):85–110.

Hjelm C-G (2006) MONA – Microdata online access at Statistics Sweden. In: Monographs of official statistics: work session on statistical data confidentiality. Publications Office of the European Communities, Luxembourg, pp 21–28.

Jackson P (2018) From ‘intruders’ to ‘partners’ – the evolution of the relationship between the research community and sources of official administrative data’. Chapter 2 in this volume.

Jenkins SP (1999) Measurement of the income distribution: an academic user’s view. In: Proceedings of the seventh CEIES seminar: income distribution and different sources of income, Cologne, Germany, 10–11 May 1999. Publications Office of the European Communities, Luxembourg, pp 75–84.

Malinvaud E (1987) Production statistique et progrès de la connaissance. In: Atti del Convegno sull’informazione statistica e i processi decisionali, Roma, 11-12 Dicembre 1986. Annali di Statistica, Serie IX, Vol, 7, Istat, Roma, pp 193-216.

OECD (2007) OECD principles and guidelines for access to research data from public funding. Organisation for Economic Co-operation and Development, Paris

OECD (2014) OECD Expert Group for International Collaboration on Data Access: final report. Organisation for Economic Co-operation and Development, Paris.

Statistics Denmark (2014) Feasibility study regarding research access to Nordic microdata. http://simsam.nu/wp-content/uploads/2016/08/Feasibility-study-regarding-research-access-to-nordic-microdata.pdf

Royal Society (2012) Science as an open enterprise. Royal Society Science Policy Centre, London.

Thaulow J, Nielsen C (2015) New Nordic model for researchers joint access to data from the Nordic statistical institutions. Paper presented at the Joint UNECE/Eurostat work session on statistical data confidentiality, Helsinki, Finland, 5–7 October 2015. http://www1.unece.org/stat/platform/display/SDCWS15/Statistical+Data+Confidentiality+Work+Session+Oct+2015+Home

Trivellato U (2000) Data access versus privacy: an analytical user's perspective. Statistica 60(4):669–689.

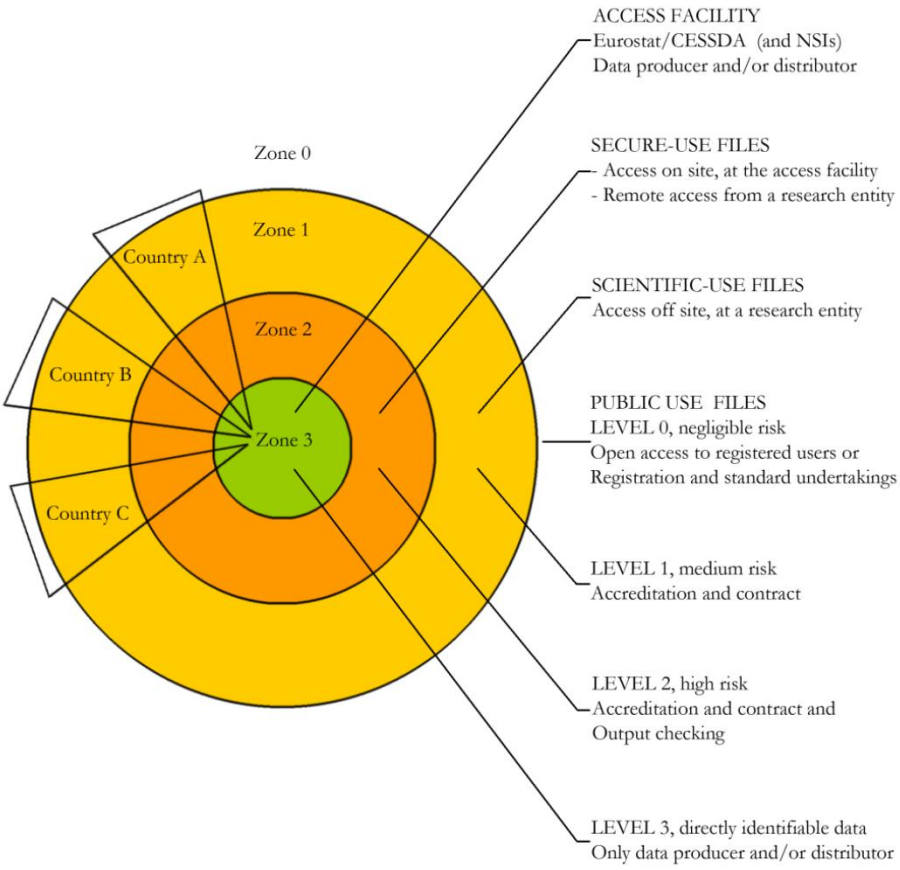
Tubaro P, Silberman R, Kleiner B et al. (2015) Researcher accreditation: current practice, essential features, and a future standard. EU Seventh Framework Programme, Project No 262608, Data without Boundaries, Deliverable 3.1. http://www.dwbproject.org/export/sites/default/about/public_deliveraples/dwb_d3-1_researchers-accreditation_report_final.pdf.

UNECE (1992) The fundamental principles of official statistics in the region of the Economic Commission for Europe. http://www.unece.org/fileadmin/DAM/stats/documents/e/ 1992/32.e.pdf.

UNECE (2007) Managing statistical confidentiality and microdata access: principles and guidelines of good practice. United Nations, New York and Geneva.

Wagner GG (1999) An economist's viewpoint of prospects and some theoretical considerations for a better cooperation: a German experience. In: Research and development: academic and official statistics cooperation. Publications Office of the European Communities, Luxembourg, pp 89–104.

Figure 1: A set of secure open environments for microdata access for research purposes*



*Adapted from OECD (2014, p. 8).