EUROPEAN COMMISSION
EUROSTAT

This paper contributes to Session 4 "Micro data: a push for transparency" of the 8th European Central Bank's Statistics Conference (6 July 2016).

Access to microdata for scientific and statistical purposes

1. INTRODUCTION

Microdata play an essential role as primary data source in the production of official statistics. Besides 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 provide the possibility to assess the underlying structure and causal links of the studied phenomena. At the same time, the calls for governments' transparency and accountability are influenced not least by the open data movement, aiming at improved availability and reuse of data. On the other hand, the society's perspective on the protection of privacy and the firm expectation of people and businesses that their personal and sensitive information are protected from disclosure create objective constraints to the potential extension of the microdata use.

The European Statistical System (ESS) has been actively supporting the use of microdata for scientific purposes, as foreseen in EU legislation. Services for the research community around accessing anonymised microdata held by the ESS, as well as their actual use, have considerably increased over the last two decades.

Moreover, in accordance with the enabling provisions in EU legislation, microdata have started to be exchanged between national statistical authorities (bilaterally or through secure data hubs in Eurostat) for statistical purposes, i.e. to improve the quality of statistics describing transnational phenomena, e.g. migration, multinational enterprises and intra-EU trade in goods. Most of those exchanges took place on an ad hoc or pilot basis. Only the provision of certain microdata regarding the structure of multinational enterprises to feed the EuroGroups' Register (EGR) is based on a legal obligation. In any case, the exchange of microdata for statistical purposes is subject to a strict data protection both in terms of confidentiality and security.

While the ESS is currently looking into the possibility to establish a regular exchange of microdata for the production of intra-EU trade in goods statistics, a thorough reflection is also needed on what more can be done to better use available microdata for research and increased transparency. The goal is to respond jointly to user expectations and grasp the opportunities created by the technological progress that allows a more efficient and profound data analysis. At the same time, protection against disclosure of confidential data as firmly established in European and national legislation and the European Statistics Code of Practice needs to be ensured in order to maintain respondents' trust in

statistical authorities' professionalism and reliability. Thus, a good operational balance between optimising the use of available data, including microdata, collected for statistical purposes and the protection of those data's confidentiality has to be found.

This paper outlines the current ESS practices regarding the access to microdata for scientific purposes and microdata exchange for statistical production purposes. The relevant data confidentiality and security aspects are also presented. Next, it reflects on the possibilities to enhance the use of microdata to better serve the society's need for a deeper understanding of the social, economic and financial developments. This outlook covers both a short term and a long-term perspective on possible new approaches to providing access to microdata while protecting privacy, taking into account the wealth of new data sources, the risk of identification with increased public availability of data and rapidly evolving technology for data linking.

2. CURRENT ESS PRACTICES REGARDING MICRODATA

2.1. Public use files

So far, microdata have been only available for statistical purposes (to produce statistics) and for scientific purposes. In the beginning of 2017 microdata from the Labour Force Survey (LFS) and EU Statistics on Income and Living Conditions (SILC) will also become available for the general public in the format of public use files (PUFs). Data records in PUFs are anonymised in such a way that the statistical units cannot be identified, either directly or indirectly.

The approach taken in the two files will be quite different. In the LFS traditional measures of statistical disclosure control will be applied, e.g. combination and deletion of breakdowns, rounding and suppression. For EU-SILC the microdata will be fully synthetic, i.e. on the basis of the original distributions artificial microdata will be generated.

2.2. Access to microdata for scientific purposes

Eurostat and the other ESS members recognise the benefits of access to (anonymised) microdata[1] by researchers. It allows multivariate analysis and cluster analysis for a better understanding of the patterns in the data and ultimately in the behaviour of social and economic actors. Also more advanced models that assess the impact of policy measures on specific groups require microdata as an input. In this way access to microdata opens up ways to understand and present official statistics far beyond the traditional use of tables and graphs and beyond the established set of domains and characteristics. Furthermore, by giving access to micro- and metadata to researchers, statistical authorities allow for independent scrutiny of their statistics, thus confirming their commitment to *scientific principles* as referred to in the UN Fundamental Principles of Official Statistics and in the European Statistics Code of Practice. By doing so, they contribute to maintaining trust in official statistics and to increasing the transparency of public administration.

[1] Set of records containing information on individual persons, households or business entities. With "microdata" we also refer to files containing confidential data. Confidential data means data allowing statistical units (individual persons, households or business entities) to be identified, thus disclosing individual information.

At the European level, Eurostat provides access to **microdata to researchers** belonging to "recognized research entities". The corresponding practices at the national level vary across the EU, e.g. in some national statistical systems individual researchers may apply for access and assignment to any organisation is not required.[2]

The microdata access systems usually offer data on persons and households. Enterprise based data are less often available. This is because (big) enterprises are in general easily identifiable even if directly identifying characteristics (name, address, business register number) are not provided. Moreover, big enterprises are usually all selected for surveys so that there is no ambiguity due to sampling. Interrelationships between businesses add additional risks. The information about an individual data provider may not be so sensitive on its own, but could lead to disclosure of data at the firms' higher organisational level.

Currently Eurostat can give access to research entities to the following micro-datasets:

- European Community Household Panel
- European Union Labour Force Survey
- Community Innovation Survey
- European Union Statistics on Income and Living Conditions
- Structure of Earnings Survey
- Adult Education Survey
- European Road Freight Transport Survey
- European Health Interview Survey
- Continuing Vocational Training Survey
- Community Statistics on Information Society
- Micro-Moments Dataset

In addition, it is planned to make the results of the Household Budget Survey available for scientific purposes still in 2016.

The procedures for accessing microdata at the European level consist of two steps:[3] recognition as a research entity and approval of the research proposal. The applicable criteria for being recognised as research entity comprise the organisation's purpose, scientific publications, independence and autonomy in formulating scientific conclusions and data protection measures put in place. The list of recognised research entities is public.[4] Once recognised, an organisation can in a second step apply for access to a specific microdata set, by describing the research proposal and justifying the need for the microdata use. The proposal is assessed by Eurostat in collaboration with the National Statistical Institutes (NSIs) in the countries that provided the data. If an NSI objects, the data of that country is removed from the microdata file which is to be made available to the recognised research entity.

---

[2] Seventh Framework Programme project Data without Boundaries (DwB) developed an exhaustive overview of microdata access systems in Europe, see: DwB Database on National Accreditation & Data Access Conditions or CIMES (Centralising and Integrating Metadata from European Statistics https://cimes.casd.eu/).

[3] Commission Regulation (EU) No 557/2013 on access to confidential data for scientific purposes. The Regulation covers the microdata available in Eurostat. In addition, further microdata sets are provided by national statistical authorities.

[4] http://ec.europa.eu/eurostat/documents/203647/771732/Recognised-research-entities.pdf

In their research proposals, the researchers indicate the mode of access which they prefer. The available options differ in terms of physical access to data and the statistical disclosure controls applied:[5]

- **Scientific use files**: data on CDs/DVDs are sent to researchers of the recognised research entities; this represents about 95% of the requests for microdata;

- **Secure use files**: researchers come to the safe centre in Eurostat (Luxembourg) to work on the data and all their final output is checked by Eurostat staff in order to ensure that there are no confidential values. This mode of access accounts for the remaining 5% of requests for microdata.

Due to the skewed distributions in business statistics, the preparation of scientific use files leads to "hiding" a significant part of the information. For this reason microdata on businesses are more useful in the format of secure use files in the Eurostat safe centre rather than scientific use files.

The increasing available computing power and the skewedness of the data distributions in business statistics raised the risk of disclosure of some of the big enterprises. In order to reduce the number of data which are hidden in statistical databases, an agreement is often sought with the entity concerned (survey respondent) on which data can be disclosed (the so called "**waiver approach**"). While this is an expensive method, as the agreements need to be reached, documented and maintained (which is not easy given the high dynamics associated with businesses over time), a selective application of this method is a powerful tool, especially for aggregate data.

2.3. Exchange of microdata for statistical purposes

In accordance with the enabling provisions in the European Statistical Law (Regulation 223/2009) and with the ESS Vision 2020,[6] the national statistical authorities in the ESS are **exchanging microdata** for statistical production purposes in specific domains. Especially for areas with a transnational dimension such as multinational enterprise groups, foreign trade statistics and international migration, the exchange of microdata offers opportunities: i.e. quality improvement by reducing bilateral asymmetries; and avoidance of double work. It requires safe data transmission tools and infrastructure as well as an appropriate access management system; both are now available or will soon be implemented in the ESS in line with the core principles for the exchange of confidential data on businesses which have been agreed by the European Statistical System Committee (ESSC) last February.

The exchange of confidential microdata for statistical purposes is organised in different ways depending on the legal bases and specific needs in the respective statistical domains. For example, microdata on multinational enterprises that feed EuroGroups' Register (EGR) are transmitted to Eurostat on a mandatory basis. The relevant experts in each national statistical institute have a secure remote access to the microdata stored in the EGR. This iterative collaboration makes it possible to maximise the coherence and reliability of information on multinational enterprise groups which is necessary for producing a number of statistics which are essential for policy analysis on economic globalisation.

---

5 http://ec.europa.eu/eurostat/web/microdata/overview.

6 http://ec.europa.eu/eurostat/web/ess/about-us/ess-vision-2020

Exchange of microdata is also a key element of ESS activities aiming at modernising intra-EU trade in goods statistics. An ESS project implementing the ESS Vision 2020 proved that exchange of microdata regarding intra-EU trade through a secure data hub in Eurostat was technically feasible and safe. Another ESS project examined the costs and benefits of several options for modernising international trade in goods statistics, in particular with regard to quality and administrative burden involved. The results of that project and the subsequent strategic discussion in the ESS showed that microdata exchange is expected to bring important value added while it is also possible to reduce considerably administrative burden (on respondents and statistical producers). The details of the future legal framework for modernising this statistical domain are being currently discussed within the ESS at the expert level.

Finally, microdata on migration were exchanged between some Member States on an ad hoc and voluntary basis with a view to clarifying and reducing bilateral asymmetries. This practice has been perceived as very useful for the participating statistical authorities.

3. TOWARDS FURTHER ENHANCING THE USE OF MICRODATA?

In the world of emerging new data sources and enhanced possibilities to integrate and analyse data from different sources, there is a naturally increasing expectation that information collected via statistical surveys and administrative registers is available to more users and for other purposes than research and statistical production. The current legal framework at the European and (even more so) national level do not leave much room for extending the use of microdata due to the strong protection required for confidential data. Nevertheless, official statisticians discuss how to better respond to user needs also in this respect.

Some improvements of access to microdata have been discussed with the partners in the ESS and would be feasible within the current legal framework, e.g. development of infrastructures and tools to allow for **tailored analysis of data** going beyond the traditional services provided by the ESS. The user should be enabled to specify himself the statistical task to be executed on microdata. Those specifications would be transmitted to a Remote Analysis Server in the statistical institute that stores the relevant microdata. The programme specifications would be executed automatically and the disclosure control of the output would be an integral part of the process ('on the fly' statistical disclosure control[7]), while the exact method and impact on the output would depend on the statistical task executed. For example, the output of a regression analysis on a sufficient high number of observations bears generally a low risk of disclosure, whereas a magnitude table has to fulfil several requirements (e.g. number of observations per cell, dominance of the largest contributor).

While the benefits of such remote analysis would be already high within the current set-up, they could be optimised with the provision by Eurostat and the statistical institutes of a so called 'data lake', i.e. an unstructured (micro-)database, allowing for the integration and analytical exploitation of multiple data sources. However, in order to establish conditions such an extended remote analysis, an agreement and possibly changes to the legal framework would be necessary.

[7] See for instance https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2013/29.pdf

Official statisticians understand very well the basic tension between the access to microdata and data confidentiality in the context of the public pressure for more transparency and more active use of microdata. Any further going initiatives to broaden or deepen the access to confidential microdata would require a review of some aspects of the current 'contract' between statistics and society regarding the use of confidential information which official statisticians collect and process. Such longer-term initiatives might include for example: full harmonisation of national practices and understanding regarding which information should be regarded as confidential (e.g. publicly available information on businesses); principle of passive confidentiality (data confidential only if justified by data subject); broader scope of transmission of microdata on businesses to Eurostat; extension of the access to confidential data to purposes similar to research (e.g. policy analysis); free publication of historical data on businesses beyond an agreed date in the past (considering that after a certain time microdata no longer disclose any sensitive market information and their disclosure can no longer have consequences for the respondent);

Such changes would require considerable time and much consensus. Proposals have to be discussed and agreed not only with the National Statistical Institutes, but also with respondents and users as the business model of official statistics hinges upon a wide understanding and sharing within society of the approaches to statistical confidentiality and access to microdata.

4. CONCLUSIONS

The European Statistical System is gradually expanding the access to confidential microdata for the production of European Statistics and for research with a view to fully exploiting its information value for scientific and statistical purposes. While it might be possible to broaden the approach by advancing with regard to the societal consensus on the areas for which microdata access is considered justified, such as policy analysis (under well specified conditions) and by relaxing the approach to statistical confidentiality in selected cases, like e.g. for historical data, such changes would require not only consensus between national statistical authorities and stakeholders, but in the most cases also substantial changes to the relevant legal provisions. The most promising and reachable development is the possibility for tailored analysis of microdata by statistical authorities acting at the request of interested users.

The future of official statistics lies with maximising the use of different data sources, involving enhanced data integration and analysis. The ESS will be prepared for such challenges in terms of technological requirements, e.g. IT security. The possibility to extend the use of microdata and enable more access to external users should be part of any strategic debate in this context.

Annex

Table 1. European legislation related to protection of personal and/or confidential data

Treaty on the functioning of the European Union - Article 338(2) of the consolidated version	*"The production of Union statistics shall conform to impartiality, reliability, objectivity, scientific independence, cost-effectiveness and statistical confidentiality; it shall not entail excessive burdens on economic operators."*
Regulation (EC) No 223/2009 on European statistics - especially Chapter V "Statistical confidentiality"	The "statistical law" defines the terms and lays down basic concepts related to statistical confidentiality
Regulation (EU) No 557/2013 on access to confidential data for scientific purposes	The Regulation establishes conditions of access to microdata
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. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)	The Directive lays down measures of personal data protection in the EU. It was transposed into national law in the EU and EEA countries. The Directive and related national legal acts will be replaced by the General Data Protection Regulation. While the Regulation entered into force on 24 May 2016, it shall apply from 25 May 2018.
Regulation (EC) No 45/2001 on the protection of individuals with regard to the processing of personal data by the Community institutions and bodies and on the free movement of such data	The Regulation implements the measures of Directive 95/46/EC in the European institutions.
European Statistics Code of Practice	The Code of Practice lays down principles covering the institutional environment, the statistical production process and the output of statistics; principle 7 is devoted to statistical confidentiality