CONFERENCE OF EUROPEAN STATISTICIANS
Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality Session 5
29-31 October 2019, the Hague, the Netherlands 08 October 2019

Data protection laws and methods in official statistics

Aleksandra Bujnowska, Wim Kloek, Fabio Ricciato (Eurostat)
Aleksandra.BUJNOWSKA@ec.europa.eu

Abstract and Paper

The European statistical law states that national statistical institutes and Eurostat shall take all necessary regulatory, administrative, technical and organisational measures to ensure physical and logical protection of confidential data. The paper positions statistical confidentiality in the range of protection measures and vis-à-vis personal data protection frameworks.

In view of exploitation of new data sources and in particular privately held personal data, Eurostat is investigating the potential of secure private computing technology and specifically the secure multiparty computation. In this contribution we briefly discuss the complementary roles of these techniques and statistical disclosure control methods.

Data protection laws and methods in official statistics

Aleksandra Bujnowska[*], Wim Kloek[**], Fabio Ricciato[***]

[*] Eurostat, aleksandra.bujnowska@ec.europa.eu
[**] Eurostat, wilhelmus.kloek@ec.europa.eu
[***] Eurostat, fabio.ricciato@ec.europa.eu

Abstract: The European statistical law (Regulation (EU) No 223/2009 on European statistics) states that national statistical institutes and Eurostat shall take all necessary regulatory, administrative, technical and organisational measures to ensure physical and logical protection of statistically confidential data. The paper positions statistical confidentiality in the range of protection measures and vis-à-vis personal data protection frameworks.
For the exploitation of new data sources, and in particular privately held data, Eurostat – in collaboration with other European Statistical System members and partners – is investigating the potential of Privacy-Preserving Computation Technologies[1]. We briefly discuss the complementary roles of these techniques with statistical disclosure control methods.

1 Objectives

The entry into force in the European Union of the measures of the General Data Protection Regulation[2] (GDPR) underlined the links between the legal frameworks for personal data protection and the frameworks for statistical confidentiality. This paper explains the relationships between both frameworks and positions statistical confidentiality, and in particular statistical disclosure control (SDC), in the range of measures available to protect data. It points out that a new framework is needed for accessing new data sources for statistical purposes.

2 Review of terms and definitions

Both legal frameworks, for statistics and for personal data protection, have their own terms and definitions. The terminologies used in these frameworks were developed independently: they might be different, but not in contradiction to each other. In order to prevent possible confusion, it is convenient to start this paper with a glossary of the main terms and concepts used in this paper along with their attribution to the legal frameworks mentioned above.

[1] http://publications.officialstatistics.org/handbooks/privacy-preserving-techniques-handbook/UN%20Handbook%20for%20Privacy-Preserving%20Techniques.pdf
[2] 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 - GDPR).

‘**Statistical framework**’ – legal framework establishing the rules and conditions for use of data for the production of official statistics.

‘**Confidential data in statistics**’ or ‘**statistically confidential data**’ –in official statistics (European statistics) ‘confidential data’ means *data which allow statistical units to be identified, either directly or indirectly, thereby disclosing individual information. To determine whether a statistical unit is identifiable, account shall be taken of all relevant means that might reasonably be used by a third party to identify the statistical unit*. (Article 3 Definitions in the Regulation 223/2009 on European statistics).

‘**Statistical confidentiality**’ - in the statistical framework - *the protection of confidential data related to single statistical units which are obtained directly for statistical purposes or indirectly from administrative or other sources and implying the prohibition of use for non-statistical purposes of the data obtained and of their unlawful disclosure* (Article 2 Statistical principles of the Regulation 223/2009 on European statistics). Statistical confidentiality cover different measures (regulatory, administrative, technical, organisational) that ensure physical and logical protection of confidential data through the whole production process.

‘**Statistical unit**’ in European statistics (Article 3 Definitions in the Regulation 223/2009 on European statistics) means the basic observation unit, namely a natural person, a household, an economic operator and other undertakings, referred to by the data.

‘**Statistical disclosure control (SDC)**’ – in European statistics - methods applied to reduce/eliminate the risk of disclosing information on the statistical units, usually based on restricting the amount of, or modifying, the data released. SDC ensures logical protection of data. SDC is linked to data dissemination and release.

‘**Anonymised data**’ – in European statistics (Regulation 223/2009 on European statistics) data prepared in such a way *that the statistical unit cannot be identified, either directly or indirectly, when account is taken of all relevant means that might reasonably be used by a third party*. The aim is to eliminate the risk of identification of statistical unit.

‘**Partially anonymised data**’ – in statistics – the data where the risk of identification is reduced, but not entirely eliminated; scientific use files for researchers are partially anonymised data.

‘**Personal data protection framework**’ – legal framework establishing the rules and conditions for use of personal data. In the EU this legal framework is GDPR.

'**Personal data**' means *any information relating to an identified or identifiable natural person ("data subject"). An identifiable person is someone who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his or her physical, physiological, mental, economic, cultural or social identity* (Article 4 Definitions of the GDPR)

'**Data subject**' in the personal data protection framework is the identified or identifiable natural person to which personal data relate (Article 4 Definitions of the GDPR).

'**Confidential data**' – broad definition – is any data that are kept secret and cannot be revealed to non-authorised persons (authorised persons are clearly defined). Confidential data in the broad sense cover: personal data, statistical confidential data, business sensitive data etc.

Business sensitive data – data that are considered a business asset and should not be disclosed to potential competitors in order to protect its legitimate business interests.

'**Encryption**' – techniques applied to make the data safe while not in use. Encryption does not ensure data anonymisation. Encryption ensures physical protection of data.

'**Disclosure**'- the situation when confidential data is accidently or maliciously revealed to non-authorized persons.

3 Statistically confidential or personal data?

In most countries, collection and processing of data on persons is regulated by laws, which specify the conditions and rules of use. These are personal data protection laws. They pose restrictions on the purpose for which the data holder may (re)use the data. Laws regulate also the rights of data subjects, and in particular guarantee that data subjects keep some control on the use of their data, in spite of the actual loss of physical control. At the European level, GDPR is a reference legal basis for personal data protection. This Regulation obliges the data holders to respect the data subjects' rights to access, change and erase the data, if so requested[3]. GDPR applies to personal data collected for all kinds of purposes: administrative, commercial, statistical etc.

"Statistical legal framework" refers to objectives, tasks and obligations of official statistical authorities. For European statistics, Regulation 223/2009 on European statistics is the reference legal framework. Whenever data protection is concerned, the statistical framework must comply with the personal data protection framework. In statistics, data protection is called "statistical confidentiality". Statistical

[3] For more information on data subjects rights, see chapter III of the GDPR.

confidentiality is a fundamental principle of **official** statistics[4]. It means that data on individual persons or business entities may be used only for statistical purposes and that appropriate measures shall be applied to prevent the disclosure of information concerning an individual person or business entity[5]. The obligations stemming from statistical confidentiality are usually stricter than personal data protection provisions.

Table 1. Comparisons of terms and definitions used in the personal data protection framework and the statistical framework in the European law

Personal data protection framework (GDPR)	**Statistical framework Regulation (EU) No 223/2009 on European statistics**
Definitions	
Personal data means any information relating to an identified or identifiable natural person ("data subject"). An identifiable person is someone who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his or her physical, physiological, mental, economic, cultural or social identity.	**'(Statistically) confidential data'** means data which allow statistical units to be identified, either directly or indirectly, thereby disclosing individual information. To determine whether a statistical unit is identifiable, account shall be taken of all relevant means that might reasonably be used by a third party to identify the statistical unit.
Data subject is the identified or identifiable natural person to which personal data relate	**'Statistical unit'** means the basic observation unit, namely a natural person, a household, an economic operator and other undertakings, referred to by the data.
Applications	
Data concerned: data on persons	**Data concerned:** data on persons, households, business entities
Entity concerned: any natural or legal person processing personal data	**Entity concerned:** members of the European Statistical System
Data processed to: provide a service to data subjects	**Data processed to:** produce statistics based on statistical unit data

Whereas the general principles and objectives of both schemes remain the same (see table 1 for comparison of major definitions in the European legal acts), the differences lie in the scope of application. Whereas GDPR applies to any personal data processed

[4] For example, principle 5 of the European Statistics Code of Practice https://ec.europa.eu/eurostat/web/products-catalogues/-/KS-02-18-142.

[5] Definition of statistical confidentiality in Article 2 of the Regulation (EU) No 223/2009 on European statistics.

by any organisation located in the EU, statistical law (Regulation 223/2009) applies to any kind of data (data based on person, households, business entities) processed by the members of the European Statistical System (ESS)[6]. Data on business entities are excluded from the scope of personal data protection framework (except for self-employed persons i.e. natural persons carrying out their own business). Personal data not used by official statistics are outside the scope of the statistical framework (see illustration on Fig 1).

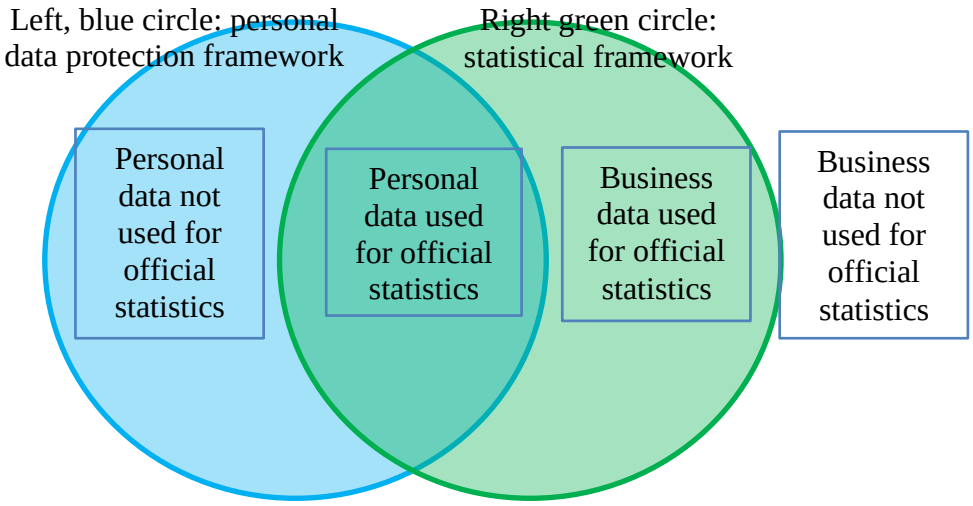

Fig 1. Application of the personal and statistical data protection framework

4 Data transmission

The data collected initially for other purposes, for example administrative, which are transmitted to statistical authorities, fall also under the umbrella of the statistical framework. From the moment the statistical authorities have access to the data the statistical framework applies. In many countries transmission of administrative data to statistical authorities is a regular, standard part of the production process. Specific laws or bilateral agreements determine the conditions of the data transfer and use[7]. The

[6] The European Statistical System (ESS) is the partnership between Eurostat and the national statistical institutes (NSIs) and other national authorities responsible in each EU country for the development, production and dissemination of European statistics. This Partnership also includes the EEA and EFTA countries. EU countries collect data and compile statistics for national and EU purposes. The ESS functions as a network in which Eurostat's role is to lead the way in the harmonization of statistics in close cooperation with the national statistical authorities.

[7] At the European level statistical offices in the EU countries transfer data to Eurostat on the regular basis. Sharing of confidential data between the statistical offices is not so usual, despite that statistical

existing regulations and long experience in keeping the administrative and survey data safe has built the trust in official statistics.

There is indeed an established framework for the transfer of data from administrative sources to statistical authorities. However this framework does not cover data originating from other sources, e.g. from private companies. The principles and practices established for administrative data might not be applicable *as is* to new data sources. A new framework needs to be established to regulate access to privately held data by statistical authorities. The new framework should take into account the multiple differences that exist between administrative data and new types of digital data, including so called "big data".

Statistical authorities are more and more interested in new data sources to complement the information already collected or to improve the quality of statistics. In order to ensure a sustainable access to such data, new legal provisions are needed to define the terms of "access" as well as to ensure a proper protection of data confidentiality in all its forms (including protection of personal data and business value) along the whole process.

The possibility to transfer personal data for statistical purposes - even if the data were initially collected for other purposes - is explicitly recognised in the GDPR. The data protection regulation allows, but does not oblige the data holders to share personal data with statistical authorities. In other words, no right of access is yet in place for statistical authorities. GDPR provides only an enabling clause that could be used as the basis for further legal or contractual arrangements for data sharing[8]. GDPR requires that, in case of personal data re-use - adequate safeguards are in place respecting the rights and freedoms of the data subjects.

5 Physical and logical protection of statistically confidential data

The European statistical authorities are obliged to take "all necessary regulatory, administrative, technical and organisational measures to ensure physical and logical (statistical disclosure control) protection of confidential data"[9].

---

law (Regulation 223/3009) explicitly recognizes this possibility under certain conditions. Exchange of confidential data is particularly important in intra-EU trade statistics. To make the data collection more efficient it was decided to use the export data collected by one country to produce statistics on import by the trading partner country. In order for the confidential data exchange to take effect the obligation on sharing the confidential data was included in Proposal for the Regulation of the European Parliament and of the Council on European business statistics (FRIBS).

[8] In case of data coming from other sources (not own data collections), the statistical authorities may be exempted from respecting specific rights of data subjects (e.g. right of access, change, erasure). Indeed, it might be very difficult for statistical authorities to follow these rights, especially if the data has been de-identified, encrypted or otherwise transformed before sharing with statistical authorities.

[9] Article 20 of the Regulation (EU) No 223/2009 on European statistics.

Physical protection refers to data security and covers the whole production process from data collection to release. Confidential data must be stored in the secure place and be accessible to authorised persons only. In most statistical organisations staff signs a confidentiality commitment, obliging them to protect confidential data and to use this data only for the production of official statistics.

Logical protection refers to the stage of data dissemination. Before publishing, all statistics must be checked for disclosure risks. Data that allow the user to derive information about a particular individual, household or business entity must be removed or modified. This process is called “statistical disclosure control" (SDC).

For statistical authorities SDC is an inherent part of the statistical process. Other data holders focus on the physical data protection as they normally use the data to provide a certain service to the data subjects; the data is not used for dissemination. For the physical protection encryption techniques are essential.

6 The case for Secure Multiparty Computation

In general, encryption techniques ensure safe storage and transfer of data. Traditional encryption techniques (symmetric and asymmetric cryptography) create a thick shell around the data to make them unusable in case of accidental or malicious take-over. However, with traditional cryptographic tools the authorised users still need to decrypt the data in order to use them. In other words, traditional cryptographic tools protect the data while “at rest” (storage) and “in transit” (transmission) but not while “in use” (processing).

Privacy-preserving computation techniques[10], such as for example secure multiparty computation (SMC), not only secure the data transfer and storage, but also allow computation of some statistics based on encrypted data. Therefore SMC can be seen a hybrid solution encompassing some aspects of both physical and logical protection of the data.

For example, SMC techniques may compute the average age of the persons without revealing the individual ages of the persons outside the domain of the data holder(s). SMC is based on the principle of *secret sharing*. Each piece of data is transformed into pieces (shares) that are sent to multiple intermediate “computing parties” (at least three). No individual party is able to reconstruct the initial data. Collusion between the computing parties must be prevented by assigning this role to entities (institutions or organisations) with guaranteed independence and confirmed credibility. By following a predetermined procedure (called “protocol” in the SMC terminology) the computing parties allow the receiver to compute exactly the average age. Along the process, no single party learns anything about the individual ages of the data subjects. Therefore,

[10] For a recent overview see the UN Handbook on Privacy-Preserving Computation Techniques https://tinyurl.com/y3rg5azm

SMC reduces the confidentiality risks of the whole data handling process. The key advantage of SMC is that it allows to receiving party (e.g. the statistical office) to learn only the final output of the computation process (the average age in this example), not the input data (the individual ages). However, SMC alone does not guarantee that the final result is not disclosive. To eliminate the risk of individual data disclosure from the final result, the receiver must still perform additional SDC checks. In the average age example, this might include an analysis of the number and characteristics of persons whose data were used. Ideally, SDC and SMC would be integrated in a single tool that blocks the computation of statistics failing to pass the disclosure check (for example requiring a minimum number of persons to calculate average). While some existing SMC platforms already incorporate some elementary SDC checks[11], the integration of SMC and SDC remains an interesting direction for future developments.

Without a solid SDC component, SMC may still be used when the receiving party is semi-trusted[12] for example between external data holders and the statistical offices, as shown in the leftmost part of Fig. 2. In this way, statistics can be computed without centralising all the input data at a single place, namely the statistical offices premises. This approach is fully in line with the principle of data minimisation in GDPR, stating that only the necessary amount of data should be processed. Also, by avoiding the concentration of large amounts of data at the premises of a single institution ("data honeypots"), SMC lowers the risks associated to intrusions and security breaches.

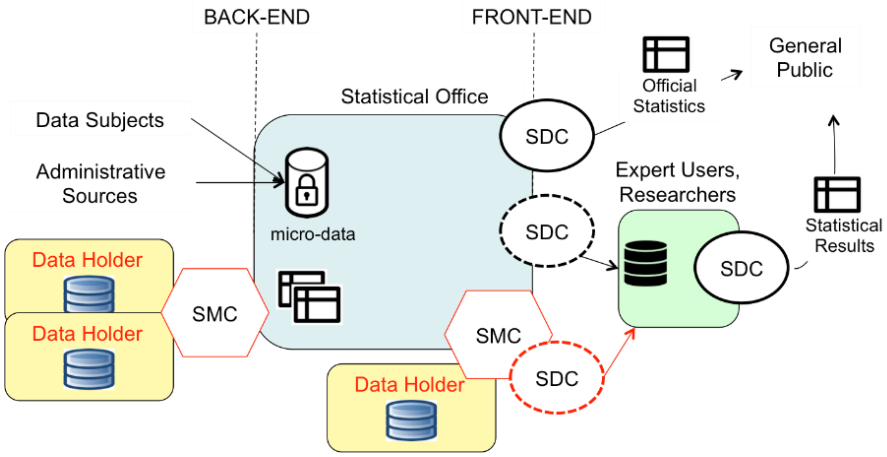

Fig 2 SMC as a technique allowing access to confidential data without transferring them

[11] See e.g. https://eprint.iacr.org/2014/512.pdf

[12] Fully trusted receiver might have access to all input and output data, while non-trusted receiver should get access to output data only after SDC check. In between these two extremes, a semi-trusted receiver is the one that can access only output data (not input) without necessarily applying disclosure checks.

7 Conclusions

The existing legal frameworks for data protection in Europe impose strong requirements on data holders. They must respect the rights of data subjects who remain in control over the way their data are used. The data holders must keep the data safe and use it for the agreed initial purpose and for strictly limited secondary purposes.

In parallel to that, statistical authorities have their own statistical legal framework that includes provisions on statistical data protection (statistical confidentiality). Physical and logical protection of statistical data is the obligation of statistical authorities. Over the years, they have implemented legal, administrative, technical and organisational measures to ensure data protection. The data entering the statistical system immediately fall under statistical confidentiality framework.

When exploring the potential of the new data sources, and specifically those held by the private sector, statistical authorities must provide evidence that adequate data protection measures are in place to minimise risk of data misuse and personal re-identification. The new privacy-preserving computation technologies that are now at sufficient maturity level, and specifically SMC, might complement the existing physical and logical measures applied by statistical authorities. By enabling *use of data without sharing them*, such technologies bear the potential to lower the barriers for statistical offices to gain access to privately held data.