UNECE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert Meeting on Statistical Data Confidentiality

1-3 December 2021, Poland

Shedding light on the legal approach to aggregate data under the GDPR & the FFDR.

Emanuela Podda (Università di Bologna)

emnauela.podda2@unibo.it

Abstract

The variety of data impacts the choice of the anonymization techniques to be applied in protecting data and grant its availability for secondary uses. From a mere statistical point of view, this choice is also influenced by another element: whether data are organized in the form of microdata or macro data, hence its structure. Datasets, indeed, may come in a variety of shapes and structures: they can be made of only numerical or categorical attributes, or usually, as a mix of numerical and categorical attributes. To this aim, semantic is key.

Acknowledging the potential impact of these differences in the data processing is of paramount importance for assessing the risks linked to the re-use of processed data in the current context of big data. Especially, when processing personal data, the re-identification risk is naturally implied.

The taxonomic analysis of the European legal framework in force for data processing and data flow (namely the General Data Protection Regulation and the Free Flow Data Regulation) reveals, at first, that the whole implant is anchored to two mutually exclusive definitions of data: personal data and non-personal data.

Ex art. 4 of the GDPR, personal data are is "any information related to an identified or identifiable natural person" while, ex art. 3 of the FFDR non/personal data is "data other than personal data" as defined in the GDPR. Hence, the nature of data is the only element to be ascertain in order to determine the applicable rules: while non-personal data can freely flow in the Digital Single Market, the circulation of personal data is lawful provided that some conditions are respected.

In this legal framework, the only reference to the structure of data comes from art. 89 and Recitals from 156 to 162 of the GDPR (introducing a series of derogation and exceptions to the data subjects rights), and from Recital 9 of the FFDR. These provisions recall the concept of aggregate data (macro data).

According to the GDPR, aggregate data is the result of personal data processing for statistical purpose (output data) and it is considered non-personal data. The Regulation defines statistical purpose as "any operation of collection and the processing of personal data necessary for statistical surveys or for the production of statistical results." According to the FFDR "specific examples of non-personal data include aggregate and anonymised datasets used for big data analytics".

The combination of these legal provisions stresses on the concept that aggregate data are non-personal data; the European Data Protection Board (EDPB) confirmed and endorsed this approach considering that, as such, aggregate data do not fall under the scope of application of the GDPR.

This approach certainly fosters innovation in the Digital Single Market: aggregate data are the main source of IoT systems, Cloud environments, Artificial Intelligence and Machine Learning applications and products. Moreover, modern information systems, rely on data aggregation to discover unusual patterns, reduce bandwidth and energy costs, or to save storage space. Hence, its advantages for the modern economies are certainly acknowledged and are out of discussion.

However, if data aggregation is performed on personal data making the output data falling out of the scope of protection provided by the GDPR, considering it as non-personal data by default may pose challenge and risks to data subjects rights. Especially, in the cases where personal data are not been obtained directly from the data subject. For these reasons this contribution aims at shed light on the legal approach to be followed when performing the risk assessment.

.

Shedding Light on the Legal Approach to Aggregate Data Under the GDPR & the FFDR

Emanuela Podda
Alma Mater Studiorum Università di Bologna, University of Luxembourg, Università degli Studi di Torino, *emanuela.podda2@unibo.it*

Abstract:
The data taxonomy designed in the General Data Protection Regulation (GDPR) and the Free Flow Data Regulation (FFDR) seems lacking legal certainty for what concerns aggregate data. As a matter of fact, the legal framework considers it as non-personal data, albeit in the non-binding parts, even with the clear awareness that some risks persist. Moreover, the literal and contextual interpretation of the two Regulations confirms that the legal framework provided for aggregate data seems applying to entities performing that kind of processing in the public interest, but even to the one processing data in the private and business one. While the data aggregation performed in the public interest by public entities is punctually regulated with even other specific laws, the one performed in the private/business interest seems to be lacking clarity and transparency.
This paper proposes a legal reasoning and argumentation on the issue, aimed at raising awareness on the importance to promote best practices on the models developed by the statistical scientific community and applied to the public sector, for avoiding data misuses and abuses in the private and business context.

1 Aggregate Data in the Taxonomy of the GDPR & FFDR

The taxonomic analysis of the European legal framework in force for data flow (namely the General Data Protection Regulation[1] and the Free Flow Data Regulation[2]) reveals that the whole implant is anchored to two mutually exclusive definitions of data: personal data and non-personal data.[3]

Hence, the nature of data is the only element to be ascertain in order to determine the applicable rules. While non-personal data can freely flow in the Digital Single Market, the circulation of personal data is lawful provided that some conditions are respected.

Apart from these two basic definitions, only few more categories of data are defined and included in one of the two categories before mentioned. According to the GDPR and the

[1] 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).
[2] Regulation (EU) 2018/1807 of the European Parliament and of the Council of 14 November 2018 on a framework for the free flow of non-personal data in the European Union (Free Flow Data Regulation).
[3] *Ex* art. 4 of the GDPR, personal data are is "any information related to an identified or identifiable natural person" while, ex art. 3 of the FFDR non/personal data is "data other than personal data" as defined in the GDPR.

FFDR[4], aggregate data is the result of personal data processing for statistical purpose (output data) and it is considered non-personal data.

Literally recalling Recital 162 of the GDPR *the statistical purpose implies that the result of processing for statistical purposes is not personal data, but aggregate data, and this result or the personal data are not used in support of measures or decisions regarding any particular natural person.*

The definition of statistical purpose ex GDPR recalls *any operation of collection and the processing of personal data necessary for statistical surveys or for the production of statistical results*. The Regulations itemize that specific examples of non-personal data include aggregate and anonymised datasets used for big data analytics.

According to this premise two considerations, strictly dependant, are fundamental:

- aggregate data are equalized to anonymized data[5], and somehow associated,

- as such data is not used in support of measures or decisions regarding any particular natural person.

Hence, it is introduced a legal presumption on the nature of the aggregate data as non-personal data, equalizing it to anonymized data, despite the state of the art confirms that they cannot be considered equals, nor both non-personal data by default. [6]

The legal presumption, introduced by the conjunction of these rules, is confirmed and endorsed by the European Data Protection Board (EDPB)[7] clarifying that, as such, aggregate data do not fall under the scope of application of the GDPR.

This approach may indeed be effective when aggregate data results as an *output* of processing non-personal data (as farm and agricultural data for example), but not in the case it represents the *output* of processing personal data.

It is of paramount importance to distinguish.

---

[4] Extensively Recital 9 of the FFDR "*The expanding Internet of Things, artificial intelligence and machine learning, represent major sources of non-personal data, for example as a result of their deployment in automated industrial production processes. Specific examples of non-personal data include aggregate and anonymised datasets used for big data analytics, data on precision farming that can help to monitor and optimise the use of pesticides and water, or data on maintenance needs for industrial machines. If technological developments make it possible to turn anonymised data into personal data, such data are to be treated as personal data, and Regulation (EU) 2016/679 is to apply accordingly."*

[5] Finck M., Pallas F., *They who must not be identified - distinguishing personal from non-personal data under the GDPR*. In International Data Privacy Law, 2020, Vol. 10, No. 1.

[6] As one of the main recent paper on the topic: Stalla-Bourdillon S., Rossi A., *Aggregation, Synthesisation and Anonymisation - A Call For A Risk-Based Assessment of Anonymisation Approaches*, in Data Protection and Privacy: Data Protection and Artificial Intelligence, CPDP Vol. 13, Hart Publishing (2021).

[7] See as an example: EDPB Opinion 10/2017, *Opinion on safeguards and derogations under Article 89 GDPR in the context of a proposal for a Regulation on integrated farm statistics*, 20 November 2017.

Several contexts are strictly dependant on aggregate data. Firstly, the pure statistical one - regulated by specific rules[8] and performed by public entities such as the National Statistical Agencies. In this case, the transparency principle grants protection to citizens' data, on the awareness of the essential role played by such processing for the society, and that, to accomplish such scope, personal data cannot be disregarded.

Moreover, aggregate data are the main source of IoT systems, Cloud environments, Artificial Intelligence and Machine Learning applications and products. Modern information systems, rely on data aggregation[9] to discover unusual patterns[10], reduce bandwidth and energy costs[11], or to save storage space[12].

Certain products and services built on aggregate data cannot even be developed without processing personal data and data cannot always be directly acquired for individuals with informed consent. This is also the *ratio* behind article 13 of the GDPR.

Often such activities are performed by private entities which might not have in place the same institutional safeguards and level of ethic as public entities.

In those case, it appears reasonable to question how large the aggregate should be before the data ceases to be "*personal*". As a consequence, questioning and investigating how to tackle the risks of this presumption, when developing ML models and AI products, even if this presumption lays in the non-binding Recitals of the Regulations[13] creating, as said, a grey area.

Academics have indeed largely demonstrated the risks linked to improper aggregation of personal data stressing, for example but not only, on the importance of a privacy preserving level of granularity. For instance, it has indeed demonstrated that is possible to infer knowledge on individuals from aggregate data[14], reason why, especially during the Pandemic the research community[15] largely promoted decentralized approaches in

[8] As a main reference: Regulation N°223/2009 on European statistics and the European Statistics Code of Practice.

[9] Cai S., Gallina B., Nyström D., Seceleanu C., *Data aggregation processes: a survey, a taxonomy, and design guidelines*, in Computing (2019) 101:1397–1429.

[10] Gray J., Chaudhuri S., Bosworth A., Layman A., Reichart D., Venkatrao M., Pellow F., Pirahesh H., *Data cube: a relational aggregation operator generalizing group-by, cross-tab, and sub-totals*. Data Mining Knowledge Discovery, 1(1):29–53, (1997).

[11] Fasolo E., Rossi M., Widmer J., Zorzi M., *In-network aggregation techniques for wireless sensor networks: a survey*. IEEE Wirel Commun 14(2):70–87, (2007).

[12] Iftikhar N, *Integration, aggregation and exchange of farming device data: a high level perspective*. In: Proceedings of the 2nd international conference on the applications of digital information and web technologies, pp 14–19, (2009).

[13] The reference to the non-binding provisions is to the Recitals of the Regulations which indeed have no binding effect, but contextualize the binding rules, providing orientation and guidance in the interpretation of the Regulation.

[14] See for example, with reference with location data: Singh R., Haasler I., Zhang Q., Karlsson J., and Chen Y.,, *Inference with Aggregate Data: An Optimal Transport Approach*, in ArXiv (2020).

[15] Nanni M., Domingo-Ferrer J., et al., *Give more data, awareness and control to individual citizens, and they will help COVID-19 containment*, in ArXiv (2020).

contact tracing apps. Moreover, more and more studies are published on Unbiased ML, fair ML[16] stressing on several aspects, and on the importance of the level of granularity in aggregate data used for AI[17] and ML. Lastly, the risks linked to ML inversion models.

To this extent, presuming that the aggregate data used to develop such models and products are non-personal data may have an impact on their fairness, not to mention the eventual risks in terms of re-identification and profiling. Lastly, the eventual risk of inversion attack on training data[18].

2 Risk Test on Aggregate Data

The legal uncertainty due to this presumption seems exacerbated when investigating what kind of safeguards are imposed by the GDPR to impact and reduce re-identification and profiling risks.

Here, two kinds of considerations should be made: one related to the risk test and consequent impact assessment proposed by the GDPR, and one related to the exceptions and derogations granted to data controllers when processing data for statistical purposes *ex* art 89 of the GDPR.

Reasoning on the Data Protection Impact Assessment (DPIA) *ex* art. 35 of the GDPR, *prima facie*, even in line with the legal implant of the FFDR, data aggregation *per se* does not seem to be a high-risk processing. On the same note, the Working Party 29 Guidelines on Data Protection Impact Assessment (DPIA) do not include such processing among the ones likely to result in a high risk for the purposes of the GDPR[19]. Hence, a DPIA seems not to be needed by data controllers.

The second consideration concerns the exceptions recalled by art. 89 and Recitals 156 to 162 of the GDPR. Article 89, indeed, provides with a series of exceptions and

---

[16] Mehrabi N., Morstatter, F., Saxena N., Lerma K., Galstyan A., *A Survey on Bias and Fairness in Machine Learning*, in ArXiv (2019).
Moreover, https://towardsdatascience.com/understanding-bias-and-fairness-in-ai-systems-6f7fbfe267f3.

[17] Meurisch, C., Mühlhäuser M., *Data Protection in AI Services: A Survey*, in ACM Computing Surveys, vol. 54, iss. 2, pg 1, (2021).

[18] Among the first papers analysing the problems of inversion attack: Golić JDj., *On the security of nonlinear filter generators*. In: Fast software encryption 1996. Lecture notes in computer science, vol 1039. Springer, Berlin, pp 173–188, (1996); Golić JDj., Clark A., Dawson E., *Generalized inversion attack on nonlinear filter generators*. IEEE Trans Comput 49(10): 1100–1108, (2000); Leveiller S., Boutros J., Guillot P., Zémor G., *Cryptanalysis of nonlinear filter generators with (0, 1)-metric Viterbi decoding*. In: Cryptography and coding – 8th IMA international conference, UK, 17–19 December 2001. Lecture notes in computer science, vol 2260. Springer, Berlin, pp 402–414, (2001); Fredrikson M., Somesh J., Ristenpart T., *Model inversion attacks that exploit confidence information and basic countermeasures*. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, pp. 1322-1333. (2015).

[19] Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679, wp248rev.01.

derogations on data subject rights, the so-called *mandated disclosures*[20], when data are processed for statistical purposes. Rights and freedoms of data subjects are restricted on the fundamental interest of carrying out such processing in the public interests[21]. But how to tackle the risks when such processing is used by private entities in the pursuing of their business interest?

Therefore, as anticipated,[22] the public statistical activity has a dedicated legal framework[23] and the rules of GDPR apply to a certain extent when processing for statistical purposes. To specify, Member States (MS) should determine statistical content, control of access, specifications for the processing of personal data for statistical purposes and appropriate measures to safeguard the rights and freedom of the data subject and for ensuring statistical confidentiality.

For example, the Italian Data Protection Authority (*Garante per la Protezione dei Dati Personali*) has issued two notes[24] introducing deontological rules regulating processing for statistical purposes, equalizing the rules to be applied when processing for statistical purposes in both public and private sector[25]. However, the Italian case might be an isolated one, because MS are free to determine such aspect in the private sector.

---

[20] Extensively on the exceptions: Ducato R., *Data protection, scientific research, and the role of information*, Computer Law & Security Review: The International Journal of Technology Law and Practice, (2020).

[21] See for example art. 179 and art. 338 on the Treaty on the Functioning of the European Union (TFEU).

[22] *Ibid.* 6.

[23] Regulation 223/2009 on European statistics. For microdata access: Regulation (EU) No 557/2013 on access to confidential data for scientific purposes. Separate laws in the EU/EEA/EFTA countries: https://ec.europa.eu/eurostat/web/ess/latest-news.

[24] Regole deontologiche per trattamenti a fini statistici o di ricerca scientifica effettuati nell'ambito del Sistema Statistico nazionale pubblicate ai sensi dell'art. 20, comma 4, del d.lgs. 10 agosto 2018, n. 101 - 19 dicembre 2018 [9069677]; Regole deontologiche per trattamenti a fini statistici o di ricerca scientifica pubblicate ai sensi dell'art. 20, comma 4, del d.lgs. 10 agosto 2018, n. 101 - 19 dicembre 2018 [9069637].

[25] For example, art. 5 defines certain criteria for the assessment of the identification risk, as follow.

1. For the purpose of communicating and disseminating statistical results, the identification risk assessment also takes into account the following criteria:

a) combinations of modalities associated with a frequency not lower than a predetermined threshold, or an intensity given by the synthesis of the values assumed by a number of statistical units equal to the aforementioned threshold, are considered aggregate data. The minimum value attributable to the threshold is three;

b) in assessing the threshold value, the level of confidentiality of the information must be taken into account;

c) statistical results relating only to public variables are not subject to the threshold rule;

d) the threshold rule may not be observed if the statistical result does not reasonably allow the identification of statistical units, having regard to the type of survey and the nature of the associated variables;

e) the statistical results relating to the same population can be disseminated in such a way that no connections are possible between them or with other known sources of information, which make possible identification;

Hence, from a mere public and institutional point of view, the main legal framework is provided by other Regulations even creating standardized mechanisms and processes for the exchange of statistical data and metadata among international organisations and their member countries (Statistical Data and Metadata eXchange)[26].

Differently, data aggregation performed in the private interest is characterized by legal uncertainty, a certain grade of fragmentation in the policies of MS and exposed to multiple risks.

Therefore, this legal uncertainty should not be neglected.

In this regard, recent research stresses[27]on the fact that a risk-based assessment for aggregation is urgently needed to tackle several risks in data protection.

Moreover, it can eventually be advisable to include the data aggregation in the DPIA as a specific tool for granting more protection to data subjects' rights and freedoms. This is especially the case when it is performed as a *further processing ex* art. 89, thus considered *de facto* compatible with the initial purpose of data collection, being outside of the scope of the compatibility test *ex* art.6 of the GDPR.

On a positive note, it should be added that risks linked data aggregation for AI products and ML models might further be addressed and regulated in the new initiatives undertaken by the European Commission in the Artificial Intelligence domain. To this extent, the analysis of the AI National Strategies[28] shows that more coordination would be needed to overcome policies fragmentation among MS, granting a minimum level of risk assessment to aggregate data.

Lastly, the author considers that the private sector should be open to contamination from the risk assessment strategies and tool developed by the institutional statistical community.

3 Conclusions and Future Work

This paper proposes a legal reasoning and argumentation on the lack of legal certainty for the risks linked to the use of aggregate data, which seems allowing that different situation (namely in the public and in the private interest) are equally considered. Moreover, it aims at bringing to the attention of the statistical community a grey area of regulation in data processing for private/business purposes.

f) confidentiality is presumed to be adequately protected if all the statistical units of a population have the same modality as a variable.

[26] See: https://sdmx.org/

[27] *Ibid*. 6.

[28] https://www.oecd.ai/dashboards/?selectedTab=countries

Future research will involve raising awareness on the importance to promote best practices for avoiding data misuses and abuses in the private sector, on the models developed by the statistical scientific community when performing data processing in the public interest.

The author considers that extending these models to the private sector can encourage data controllers to behave in a data protection friendly manner, on the awareness that comprehensive protection requires more interdisciplinary research and a combination of approaches at different levels.

References

Cai, S., Gallina, B., Nyström, D., Seceleanu, C., *Data aggregation processes: a survey, a taxonomy, and design guidelines, in Computing* (2019) 101:1397–1429.

Ducato, R., *Data protection, scientific research, and the role of information*, in Computer Law & Security Review: The International Journal of Technology Law and Practice, 37 (2020).

Fasolo, E., Rossi, M., Widmer, J., Zorzi, M., *In-network aggregation techniques for wireless sensor networks: a survey*, IEEE Wirel Commun 14(2):70–87, (2007).

Finck, M., Pallas, F., *They who must not be identi-fied—distinguishing personal from non-personal data under the GDPR.* In International Data Privacy Law, 2020, Vol. 10, No. 1.

Fredrikson M., Somesh J., Ristenpart T., *Model inversion attacks that exploit confidence information and basic countermeasures*. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, pp. 1322-1333. (2015).

Golić JDj., *On the security of nonlinear filter generators*. In: Fast software encryption 1996. Lecture notes in computer science, vol 1039. Springer, Berlin, pp 173–188, (1996).

Golić JDj., Clark A., Dawson E., *Generalized inversion attack on nonlinear filter generators*. IEEE Trans Comput 49(10): 1100–1108, (2000).

Gray, J., Chaudhuri, S., Bosworth, A., Layman, A., Reichart, D., Venkatrao, M., Pellow, F., Pirahesh, H., *Data cube: a relational aggregation operator generalizing group-by, cross-tab, and sub-totals*. Data Mining Knowledge Discovery, 1(1):29–53, (1997).

Iftikhar, N., *Integration, aggregation and exchange of farming device data: a high level perspective*. In: Proceedings of the 2nd international conference on the applications of digital information and web technologies, pp 14–19, (2009).

Leveiller S., Boutros J., Guillot P., Zémor G., *Cryptanalysis of nonlinear filter generators with (0, 1)-metric Viterbi decoding.* In: Cryptography and coding – 8th IMA international conference, UK, 17–19 December 2001. Lecture notes in computer science, vol 2260. Springer, Berlin, pp 402–414, (2001).

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A., *A Survey on Bias and Fairness in Machine Learning*, in ArXiv (2019).
Meurisch, C., Mühlhäuser M., *Data Protection in AI Services: A Survey*, in ACM Computing Surveys, vol. 54, iss. 2, pg 1, (2021).
Nanni, M., Domingo-Ferrer, J., et al., *Give more data, awareness and control to individual citizens, and they will help COVID-19 containment*, in ArXiv (2020).
Singh, R., Haasler, I., Zhang, Q., Karlsson, J., Chen, Y., *Inference with Aggregate Data: An Optimal Transport Approach*, in ArXiv (2020).
Stalla-Bourdillon S., Rossi A., *Aggregation, Synthesisation and Anonymisation - A Call For A Risk-Based Assessment of Anonymisation Approaches, in Data Protection and Privacy: Data Protection and Articficial Intelligence*, CPDP Vol. 13, Hart Publishing (2021).

Grey literature

- 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).
- Regulation (EU) 2018/1807 of the European Parliament and of the Council of 14 November 2018 on a framework for the free flow of non-personal data in the European Union (Free Flow Data Regulation).

Acknowledgments

Support from the European Commission (Horizon 2020 research & innovation programme under the Marie Sklowdoska-Curie grant agreemant No. 814177) under the coordination and supervision of Prof. Monica Palmirani, Alma Mater Studiorum Università di Bologna.