Inferring the Meaning of Non-personal, Anonymized, and Anonymous Data

Inferring the Meaning of Non-personal, Anonymized, and Anonymous Data

Emanuela Podda[1,2,3] and Monica Palmirani[1]

[1] ALMA AI, Alma Mater Studiorum, Università di Bologna, Bologna, Italy
emanuela.podda@studio.unibo.it, {emanuela.podda2, monica.palmirani}@unibo.it
[2] Università di Torino, Turin, Italy
[3] University of Luxembourg, Luxembourg, Luxembourg

Abstract. On the awareness of the dynamism pertaining to data and its processing, this paper investigates the problem of having two mutually exclusive definitions of personal and non-personal data in the legal framework in force. The taxonomic analysis of key terms and their context of application highlights the risk to crystalize the whole system upon which the digital single market is built, suffocating its future development. With this premise, the paper discusses the extent of the two main data processing tools provided by the GDPR, questioning the *ex-ante* categorization of data and its outcome, supporting stakeholders in overcoming this issue.

Keywords: Non-personal data · Anonymization · Pseudonymization

1 Introduction

Everyday people generate data from all spheres of their life, circulating in the IoT environments and *feeding* Big Data Systems (Perera 2015), posing uncountable challenges to data protection and privacy (Sollins 2019). In the last year, the European Commission proposed important initiatives to unlock the re-use of different types of data and create a common European data space[1]. The first pillars were proposed in February 2020 as the Data Strategy[2] and the White Paper on Artificial Intelligence[3], followed by the adoption of a proposed Regulation on Data Governance[4] in November 2020. Lastly, the Artificial

[1] https://ec.europa.eu/digital-single-market/en/data-policies-and-legislation-timeline.

[2] A European strategy for data, Brussels, 19.2.2020, COM(2020) 66 final.

[3] The European Commission confirms that data and artificial intelligence (AI) can help find solutions to many of society's problems, from health to farming, from security to manufacturing. However, it also stresses on the risks posed by AI. It stresses on the need to enforce it adequately to address the risks that AI systems create.

[4] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0767&from=EN

Intelligence Act[5]. In this proposal, the European Commission confirms the methodological roadmap based on risk analysis, intermediary services - already proposed in the Digital Services Act[6] - and certifications for the Artificial Intelligence processes and products.

The whole new system relies on principles and rules introduced by the two main Regulations for granting the data flow in the digital single market: the General Data Protection Regulation[7] (GDPR), and the Free Flow Data Regulation[8] (FFDR). While personal data can flow provided that some conditions are respected (e.g., consent, processing, risk evaluation, etc.) non-personal data can freely flow in the digital environment. Thus, the whole legal system is anchored to the dichotomy *personal & non-personal data,* and even its development is strictly dependent on it, facing the risk of suffocating innovation.

Since the entry into force of this legal framework, few areas of improvement have been identified[9], even on the awareness that the context and the infrastructure are rapidly evolving and changing, therefore potentials and risks. The EU Member States will set up a new common digital platform[10], the Data Space, where the international dimension will play a central role[11], creating a level playing field with companies established outside the EU[12]. The ability for private and public sector actors to collect and process data on a large scale will increase: devices, sensors and networks will create not only large volumes of data, but also new types of data like inferred, derived and aggregate data (Abuosba 2015) or synthetic data (Platzer 2021), moving beyond the data dichotomy imposed by the legal framework.

The awareness of the legal vulnerabilities pertaining to this technological evolution, implies that both law and technology must, together, promote and reinforce the beneficent use of Big Data for public good (Lane et al. 2014), but also, people's control of their personal data, their privacy and digital identity (Karen 2019)[13].

---

[5] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52021PC0206&from=EN.

[6] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0825&from=en.

[7] 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.

[8] 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.

[9] To this extent, refer to the first Report on the evaluation of the GDPR published by the Commission on June 2020 https://ec.europa.eu/info/sites/info/files/1_en_act_part1_v6_1.pdf.

[10] See the program Gaia-X, https://www.data-infra-strucure.eu/GAIAX/Navigation/EN/Home/home.html.

[11] The report stresses on the fact that "the Commission will continue to focus on promoting convergence of data protection rules as a way to ensure safe data flows".

[12] Ibid, 10.

[13] The author affirms that a sustainable IoT Big Data management can be effectively designed only after decomposing the set of drivers and objectives for security/privacy of data as well as innovation into: 1) the regulatory and social policy context; 2) economic and business context; and 3) technology and design context. By identifying these distinct objectives for the design of IoT Big Data management, a more effective design and control is possible.

2 Personal Data, Non-personal Data, Mixed Datasets

The GDPR and the FFDR provide the taxonomy of data. Art. 4(1) of the GDPR specifies that "personal data" means "*any information relating to an identified or identifiable natural person ('data subject') […]*".[14] Art. 3 of the FFDR defines "non-personal data" as data other than personal data as defined in point (1) of Article 4 of Regulation (EU) 2016/679.

These definitions are mutually exclusive and strongly chained: the definition of non-personal data is dependent on the definition of personal data. *Ex ante,* they seem not considering the *physiological attitude of data to be treated* thus, not considering the data lifecycle (Wing 2019). *De facto*, the legal framework does not provide any concrete tool able to ensure to check the nature of data during its physiological lifecycle. On the contrary, imposes to data controllers and data processors to keep monitoring the risks linked to such processing.

Data processing, indeed, modifies the status of data, its definition and category. Hence, it is necessary to distinguish between the *static perspective*, based on the reasoning of what can be literally considered as "personal data" and the *dynamic* one, specifically on what kind of *status* modification data can have due to its lifecycle.

In line with these premises, if considering both perspectives, the span of the concepts increases, proportionally implying the risk of overlapping in definitions and sclerotizing the whole system of data flow in the digital single market. Not to mention that, since these definitions are strictly dependant, any vulnerability in one affects the other and *vice-versa.*

These critical points were originally highlighted in the Impact Assessment of the Regulation[15]. Nowadays, after few years of the entry into force, they are largely confirmed and still discussed in the academic debate (Graef et al. 2018; Hu et al. 2017; Finck and Pallas 2020; Leenes 2008; Stalla-Bourdillon and Knight 2017).

The definition of personal data is coming from the centerpiece of EU legislation on data protection, Directive 95/46/EC, adopted in 1995[16] and it has been transposed in the GDPR. Since then, it led to some diversity in the practical application. For example, the issue of objects and items ("things" – referring to IoT systems) linked to individuals, such as IP addresses, unique RFID numbers, digital pictures, geo-location data and telephone numbers, has been dealt differently among Member States[17]. The CJEU played - and keeps playing - an essential role in resolving these diversities, harmonizing the legislation[18].

[14] In order to clarify the concept, the WP29 04/2007 on the concept of Personal Data states that the contextual presence of 4 elements connotes personal data: 1) Any information, 2) Relating to, 3) An identified or Identifiable, 4) Natural Person.

[15] Commission Staff Working Paper, Brussels, 25.1.2012, SEC(2012) 72 final, Impact Assessment.

[16] The Directive was also complemented by several instruments providing specific data protection rules in the area of police and judicial cooperation in criminal matters (*ex* third pillar), including Framework Decision 2008/977/JHA.

[17] These diverities are extensively treated in the Impact Assessment.

[18] To this aim, as an example, the judgment in Case C-582/14: Patrick Breyer v Bundesrepublik Deutschland.

The core of the problem leading to legal uncertainty as a major area of divergence in the Member States, and strictly linked to the data processing[19] - is related to the concept of *identifiability*. Specifically, to the circumstances in which data subjects can be said to be "***identifiable***".

The importance of this concept is strengthen by the combined provisions of Recital 26 of GDPR and Recital 8 of the FFDR were it is clearly stated that data processing can modify the nature of data. This problem acquires even more resonance, when literally recalling Art. 2(2) of the FFDR "*In the case of a data set composed of both personal and non-personal data, this Regulation applies to the non-personal data part of the data set. Where personal and non-personal data in a data set are inextricably linked, this Regulation shall not prejudice the application of Regulation (EU) 2016/679.*"

In order to clarify the concept of *inextricability*, the European Commission released a *Practical guidance for businesses on how to process mixed datasets*[20] contextualizing the case and confirming that in most real-life situations, a dataset is however very likely to be composed of both personal and non-personal data (*mixed dataset*), thus it would be challenging and impractical, if not impossible, to split such mixed dataset.

As pointed by some authors (Greaf 2018), this data taxonomy becomes counterproductive to data innovation[21].

Therefore, still nowadays, the meaning and interpretation of identifiability yet represents the main reason why the concept of personal data and its interconnection with non-personal data is widening being still problematic, especially in perspective of the data processing, e.g. anonymization, pseudonymization.

When transposed in the technological environment, this perspective leads to the concept Personally Identifiable Information (hereinafter referred as PII). Referring to

[19] Specifically, on the nature of processed data, Data Protection Authorities (hereinafter referred as DPAs) considered encoded or pseudonymised data as identifiable thus, as such, as personal data in relation to the actors who have means (the "key") for re-identifying the data, but not in relation to other persons or entities (e.g. Austria, Germany, Greece, Ireland, Luxembourg, Netherlands, Portugal, UK). In other Member States all data which can be linked to an individual were regarded as "personal", even if the data are processed by someone who has no means for such re-identification (e.g. Denmark, Finland, France, Italy, Spain, Sweden). DPAs in those Member States are "generally less demanding" with regard to the processing of data that are not immediately identifiable, taking into account the likelihood of the data subject being identified as well as the nature of the data.

[20] Guidance on the Regulation on a framework for the free flow of non-personal data in the European Union Brussels, 29.5.2019 COM(2019) 250 final.

[21] A timid tentative of overcoming this problem, it is contained in the proposal of the Data Governance Act where the Commission proposes to create a formal expert group, the European Data Innovator Boards.

the International Standards[22] ISO 27701[23] defines PII as "*any information that (1) can be used to establish a link between the information and the natural person to whom such information relates, or (2) is or can be directly or indirectly linked to a natural person*".

If considered the amount of data that can be freely gathered in the digital info-sphere and the potential of data mining tools (Clifton 2002) contextualizing these definitions in several datasets, any kind of *value* linked to a person may lead to a PII. Consequently, it could be possible to affirm that in the digital context, affected by the process of datafication (Palmirani and Martoni 2019), an *identity* is any subset of attributed values of an individual person and, therefore, usually there is no such thing as "the identity", but several of them, as many as the number of the values combined with the same *data-subject* (Pfitzmann and Hansen 2010).

The problem presents a broader span if recalling the main premise on the data cycle, thus taking into account that even any PII has a natural lifecycle (Wing 2019; Abuosba 2015). As specifically stated in the ISO standard "*from creation and origination through storage, processing, use and transmission to its eventual destruction or decay. The risks to PII can vary during its lifetime but protection of PII remains important to some extent at all stages. PII protection requirements need to be taken into account as existing and new information systems are managed through their lifecycle"*.

To this extent, it can certainly be said that what defined at the moment of *ex-ante processing* as personal data, cannot necessarily last and be confirmed at the moment of *ex-post processing* as non-personal and *vice-versa.* In this regard, *domino effect* is spilling over the definition of anonymous data. There is indeed no doubt that this category includes data, not linkable by any mean to a data subject[24] but, for which this certainty is not undoubtable as the one namely recalled by Recital 26 of the GDPR referring to *data rendered anonymous in such a way that the data subject is no longer identifiable.*

To what extent a data processing can grant that a data subject is no longer identifiable?

Academics are currently stressing on a more proper evaluation of the differential element between personal and non-personal data (Finck and Pallas 2020), and on the importance of the paramount importance of the legal Principle of Data Minimization to overcome this legal emapasse (Biega et al. 2020).

Others (Stalla-Bourdillon and Knight 2018), also referring to the Breyer case, considers that characterizing the data should be context-dependent.

---

[22] The Commission's policy aims to align European Standards as much as possible with the international standards adopted by the recognized International Standardization Organizations ISO, IEC and ITU. This process is called "primacy of international standardization", meaning that European standards should be based on International standards (COM(2011)-311, point 7). For more info, cfr: https://ec.europa.eu/growth/single-market/european-standards/policy/international-activities_en.

[23] ISO/IEC 27701:2019 (formerly known as ISO/IEC 27552 during the drafting period) is a privacy extension to ISO/IEC 27001. The design goal is to enhance the existing Information Security Management System (ISMS) with additional requirements in order to establish, implement, maintain, and continually improve a Privacy Information Management System (PIMS). The standard outlines a framework for Personally Identifiable Information (PII) Controllers and PII Processors to manage privacy controls to reduce the risk to the privacy rights of individuals.

[24] For example, those referred to businesses, those referred to industrial machinery, stars data like the ones related to Mars, labs data on chemical reactions, etc.

For others (Purtova 2018), the broad notion of personal data is not problematic and even welcome but this will change in future when everything will be personal or will contain personal data, leading to the application of data protection to everything. This will happen because technology is rapidly moving towards perfect identifiability of information, where *datafication* and data analytics will generate a lot of information.

Hence, in order to mitigate the gross risk of re-identification, contextual checks become essential and they should be conceived as complementary to sanitization techniques (Gellert 2018).

3 Anonymization Techniques in the Light of WP29 05/2014 and Its State of the Art

Anonymizing personal data implies a data processing which makes uncertain the attribution of that data to a certain person (data subject), relying on the probability calculation. Stemming from the expansion of data products usually provided by National Statistic, anonymization is considered by the Working Party 29, on the Opinion 05/2014, as a "further processing"[25]. International Standard ISO 29100 considers anonymization as the "*process by which Personally Identifiable Information (PII) is irreversibly altered in such a way that a PII principal cannot longer be identified directly or indirectly, either by the PII controller alone or in collaboration with any other party*"[26].

Differently from pseudonymity[27] which is *generally* (Mourby et al. 2018) distinguished by reversibility[28] (reason why the GDPR considers pseudonymized data still personal data) anonymization therefore should generally imply an *irreversible* alteration of personal data. The European legislation does not provide an explicit regulation on anonymization or an identification on its techniques, neither how the process should be, or could be performed. The legal focus is not on the tool *per se*, rather on its outcome.

[25] As such, must comply with the test of compatibility in accordance with the guidelines provided by the Working Party 29 Opinion 03/2013 on purpose limitation and with the de-anonymization risk test as for the Working Party 29 Opinion 05/2014.

[26] International Standard Organization (ISO/IEC) 29100:2011 Information technology – Security techniques – Privacy framework (*Technologies de l'information – Techniques de sécurité – Cadre privé*).

[27] According to Art. 4(5) GDPR 'pseudonymization' means "the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organizational measures to ensure that the personal data are not attributed to an identified or identifiable natural person".

[28] Pseudonymization is a de-identification process referenced in the GDPR as both security and data protection by design mechanism. There are different levels and scenarios of pseudonymity but as for anonymization process, different levels of security. See in details: https://www.enisa.europa.eu/publications/pseudonymisation-techniques-and-best-practices.

It is solely considered the potential risk linked to this data treatment, thus providing guidance and clarification with the Working Party 29 Opinion 05/2014[29] which, has no binding character, but by some authors (El Emam, Álvarez 2015) it is even considered lacking in some critical topics.

According to the definition provided by the Recital 26 of Directive 95/46/EC, recalled by the Opinion 05/2014, anonymization means *stripping data of sufficient elements such that the data subject can no longer be identified*. Therefore, data must be processed in such a way that it can no longer be possible to identify a natural person by using "*all the means likely reasonably to be used*" by either the controller or a third party. Such processing must be irreversible but, here again, the question to what extent this irreversibility can be granted.

There is no doubt that anonymization has a high degree of uncontrollability, but even that technological development has reached a point, as anticipated, of questioning whether anonymization can still be considered as an irreversible data processing. Moreover, the same approach seems confirmed in Recital 9 of the FFDR "*If technological developments make it possible to turn anonymized data into personal data, such data are to be treated as personal data, and Regulation (EU) 2016/679 is to apply accordingly*".

Moreover, as said, the WP29 focuses only on the outcome of anonymization strictly related to the risk of de-anonymization, elaborating only on the robustness of few technique based on three criteria:

– *is it still possible to single out an individual*[30],
– *is it still possible to link records relating to an individual*[31], and
– *can information be inferred concerning an individual*[32].

The WP29 recalls the two main anonymization techniques: randomization and generalization. Randomization alters the veracity of data weakening the links between values and objects (data subject), introducing a casual element in the data. This result can be concretely accomplished with few techniques: permutation, noise addition and differential privacy.

---

[29] The Article 29 Working Party (today EDPB – European Data Protection Board) was set up under the 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. It provides the European Commission with independent advice on data protection matters and helps in the development of harmonized policies for data protection in the EU Member States. One of the main tasks of the Article 29 WP was to adopt Opinions without a binding character but fundamental in order to clarify critical data protection issues.

[30] "The possibility to isolate some or all records which identify an individual in the dataset" WP29 Opinion 05/2014 on Anonymization Techniques, WP216, (0829/14/ EN). (2014).

[31] "The ability to link, at least, two records concerning the same data subject or a group of data subjects" WP29 Opinion 05/2014 on Anonymization Techniques, WP216, (0829/14/ EN). (2014).

[32] "The possibility to deduce, with significant probability, the value of an attribute from the values of a set of other attributes" WP29 Opinion 05/2014 on Anonymization Techniques, WP216, (0829/14/ EN). (2014).

According to the Opinion 05/2014, with differential privacy (Dinur and Kobbi 2003; Dwork 2011) singling out, inference and linkability may not represent a risk. However, statistical academics have just underlined its vulnerability (Domingo-Ferrer et al. 2021).

Differently from randomization, generalization dilutes the attributes by modifying the respective scale or order of magnitude and it can be performed using the following techniques: aggregation and K-anonymity (Samarati and Sweeney 1998) (which has been implemented with several algorithms) (Samarati 2001; Le Fevre et al. 2005; Xu 2006), L-diversity (Machanavajjhala et al. 2007) (which seems to be vulnerable and subject to probabilistic inference attacks) and T-closeness (Li et al. 2007) (as a refinement of L-diversity).

Certainly, the state of the art linked to the techniques listed in the Opinion 05/2014 seems confirming that anonymization methods face big challenges with real data and that it cannot longer be considered from a static perspective, but only from the dynamic one, being a dynamic checked process.

The evolution of the academic debate seems confirming the vulnerability of anonymization. Some academics (Ohm 2010; Nissembaum 2011; Sweeney 2001) stress on the unfeasibility of granting a proper and irreversible anonymization and at the same time maintain the data useful, or vice-versa. Others, (Cavoukian 2010; Yakowitz 2011) consider that, despite the awareness of the de-anonymization issue, a compromise between the commercial, social value of sharing data and some risks of identifying people should always be reached, even if producing consequences for personal privacy and data protection.

Moreover, moving beyond the general approach of questioning the concept of anonymization, its values and paradigm, in the last two decades the debate changes perspective. Currently, more and more authors are gathering empirical evidences on the possibility to reverse the process of anonymization, exploring and studying its correlated techniques. The attention is focused on the concrete possibility of de-anonymize data which have undergone a process of anonymization (no matter on which anonymization techniques used) due to the available technology and the technological development. Based on these assumptions, it is implicitly recognized that - within the context of the modern technology and due to uncontrollable technological development - the simple model of anonymization is unrealistic and researchers are currently exploring new models of anonymization.

For these reasons, the new trend is to combine many techniques in a pipeline using a complex monitored process, capable to provide also a dashboard where the human expert is maintained in the loop (Jakob et al. 2020).

In addition, it can be mentioned the model of "*functional anonymization*" which is based on the relationship between data and environment within which the data exists, the so-called "data environment" (Elliot et al. 2016; Elliot and Domingo Ferrer 2018). Researchers provide a formulation for describing the relationship between the data and its environment that links the legal notion of personal data with the statistical notion of disclosure control (Elliot et al. 2018; Hundepool and Willenborg 1996; Sweeney 2001, 2001b; Domingo-Ferrer and Montes 2018).

Assuming that *perfect* anonymization has failed and it is strictly linked to the context, some academics (Rubinstein and Hartzog 2016) remark that while the debate on de-anonymization remains vigorous and productive, "*there is no clear definition for policy*", arguing that the best way to move data release policy is focusing on the process of minimizing risk of re-identification and sensitive attributes disclosure, rather than trying to prevent harm.

As anticipated, traditional anonymization methods which were originally tailored for the statistical context face big challenges with real data. From a mere legal point of view, the guidance provided by the WP29 in the Opinion 05/2014 needs to be reviewed, in line with the technological development. To confirm it, the fact that recently the European Parliament has recenty adopted a resolution inviting the European Data Protection Board "*to review WP29 05/2014 of 10 April 2014 on Anonymisation Techniques"*[33].

4 …and Pseudonymization?

The Opinion 05/2014 defines pseudonymization by negation, as "*not a method of anonymization* […]. *It merely reduces the linkability of a dataset with the original identity of a data subject, and is accordingly a useful security measure.*"

The concept of pseudonymity[34] has a long history and in literature: many writers had a pseudonymous. Nowadays, the term is mostly used with regard to identity and the Internet, and ISO 25237[35] defines pseudonymization as a "*particular type of de-identification that both removes the association with a data subject, and adds an association between a particular set of characteristics relating to the data subject and one or more pseudonyms*".

The definition provided in the main legal framework in force, is slightly different, and it is contained in art. 4(5) of the GDPR: "*the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organizational measures to ensure that the personal data are not attributed to an identified or identifiable natural person*".

Despite the different perspective, despite the fact that the GDPR stresses more on the "*local linkability*" (Hu et al. 2017) there are two common elements in the definitions:

- the removal of the attribution link between the personal data and the data subject
- its replacement with new additional information.

As for anonymization, even for pseudonymization, the GDPR does not define techniques and tools, but provides orientation in terms of context. It places it in two different articles: in art. 25 recalling it as appropriate technical and organizational measure

[33] European Parliament resolution of 25 March 2021 (2020/2717(RSP)).

[34] The term pseudonymous stems from the Greek word "*ψευδώνυμον (pseudṓnymon)*" literally "false name", from *ψεῦδος (pseûdos)*, "lie, falsehood" and *ὄνομα (ónoma)*, "name".

[35] ISO 25237:2017 Health informatics—Pseudonymization. It contains principles and requirements for privacy protection using pseudonymization services for the protection of personal health information.

designed to implement data-protection principles[36], as well as in art. 32 listing it - with encryption - as a security measure that should be implemented by the data controller and the data processor.

These specific collocations explicitly confirm not only that pseudonymization represents a data security measure, but also that the tool can be implemented and adapted to the specific needs and aims of the data controller and the data processor (Drag 2018) in line with the principles of privacy by design (Cavoukian 2010).

The main reference on pseudonymization techniques stemming from the European Institutions, apart from samples recalled in the WP29 Opinions, it is provided by ENISA, the European Agency for Cybersecurity. Listing it among its priorities of the Programming Document 2018–2020, it provides recommendations on shaping technology according to GDPR provisions. Specifically, a complete guidance can be found in three recommendations[37,38,39] thus, as such, not legally binding, confirming the same approach followed by the WP29 Opinion 05/2014 on the Anonymization Techniques.

In ENISA Recommendations, different techniques are described, on the assumption that pseudonymization can relate to a single identifier, but even to more than one. The pseudonymization can be performed with the following techniques: Counter, Random Number Generator (RNG), Cryptographic Hash Function, Message Authentication Code (MAC), and Encryption.

However, not all the pseudonymization techniques are equally effective and the possible practices vary: they can be based on the basic scrambling of identifiers, or to advances cryptographic mechanism. The level of protection may vary accordingly.

In any case, especially for the hash function there is doubt to what extent it represents an efficient pseudonymization technique, especially under certain circumstances such as the case in which the original message has been deleted, thus granting irreversibility. In this case indeed, the hash value might even be considered as anonymized[40], on the basis of the dichotomy *reversible/irreversible* processing.

In term of policy, this decision is of paramount importance to determine the compliance of the rights recognized by the GDPR for certain types of processing (e.g. research,

---

[36] Specifically, art. 25(1) says that "Taking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, the controller shall, both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organizational measures, such as pseudonymization, which are designed to implement data-protection principles, such as data minimization, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects".

[37] ENISA, Recommendations on shaping technology according to GDPR provisions. An overview on data pseudonymization, November 2018.

[38] ENISA, Pseudonymization techniques and best practices. Recommendations on shaping technology according to data protection and privacy provisions, November 2019.

[39] ENISA, Data Pseudonymisation: Advanced Techniques & Use Cases, January 2021.

[40] AEPD, Introduction to the hash function as a personal data pseudonymization technique, October 2019.

traffic data analysis, geolocation, blockchain and others). The last ENISA report on January 2021[41] describes advanced techniques at the state of the art (e.g., zero-knowledge proof), demonstrating that the pseudonymization, like the anonymization, is a dynamic concept depending to the evolution of the technological over time. Additionally, this report remarks also how these techniques are very context dependent and they requires a detailed analysis of all the lifecycle of the data management including custody, key-ring management. In particular the data custodianship (or similar concepts such as data trustees or intermediaries) as a particular agent, trusted intermediaries for supporting confidentiality and protection of data. This may allow to pseudonymize the data and make them available for researchers, or can even be used in the healthcare sector.

The data custodian, or intermediary as defined in the first draft of the Data Governance Act, provides also the service to release synthetic data "*that is not directly related to the identifying data or the pseudonymised data but, still, shows sufficient structural equivalence with the original data set or share essential properties or patterns of those data. Synthetic data is being used instead of real data as training data for algorithms or for validating mathematical models.*"

The traditional research debate on pseudonymity tried to clearly define the difference between anonymization and pseudonymization, focusing on the semantic (Pfitzmann and Hansen 2010). After being included in the GDPR as a data processing tool and as a data security measure, a primary focus is given to its risks (Stevens 2017; Bolognini and Bistolfi 2017) and the ambiguity surrounding the concept of pseudonymization in the GDPR (Mourby et al. 2018).

Overall, the state of the art seems confirming that pseudonymization has a greater potential of data protection than anonymization, and the implementation of the different techniques in currently ongoing.

5 Conclusions

The legal uncertainty pertaining to the two mutually exclusive definition of personal and non-personal data is spilling over the two main data processing tools provided by the legal framework in force, and especially the anonymization one.

The current evolution of the techniques in this sector suggests to approach the problem from a dynamic perspective, using a concept of permanent lifecycle checking. This will allow a constant revision of the admissible parameters and techniques, according to the state of the art. In this respect, the proposal for the Data Governance Act seems relying on intermediary certified and trusted services, aiming to different goals: i) a correct implementation of the pseudonymization/anonymization at the best of the state of the art and case-by-case according to the context of application (e.g., health); ii) a constant risk assessment; iii) the peculiar role of data custodian capable to provide a proxy access to other third parties (e.g., research institutions) also through synthetic datasets.

According to these premises, the two mutually exclusive definitions of personal and non-personal data seem obsolete and should be revised in favor of a constant and dynamic process which uses risk analysis, supported by intermediary certified actors. Also relevant for the evolution of anonymization, the concept and role played by "data altruism

[41] ENISA, 2021.

organizations[42]" included in the Data Governance Act. Therefore, it could determine a proxy where to anonymize the data using particular conditions and techniques, thanks to the special regime and regulation of this particular processing.

Finally, because some of these anonymization techniques use artificial intelligence artifact, is also relevant the Artificial Intelligence Act which proposes, again, a more detailed risk management approach and the introduction of a European Certification (CE) of the Artificial Intelligence production processes, with related certified actors playing the role of independent intermediators, ensuring the proper application of the regulation according to technological benchmarking.

References

Abuosba, K.: Formalizing big data processing lifecycles: acquisition, serialization, aggregation, analysis, mining, knowledge representation, and information dissemination. In: 2015 International Conference and Workshop on Computing and Communication, IEMCON (2015)

Aggarwal, C.: On k-anonymity and the curse of dimensionality. In: VLDB (2005)

Biega, A.J., Potash, P., Daumé III, H., Diaz, F., Finck, M.: Operationalizing the legal principle of data minimization for personalization, computers and society. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (2020)

Bolognini, L., Bistolfi, C.: Pseudonymization and impacts of Big (personal/anonymous) Data processing in the transition from the Directive 95/46/EC to the new EU general data protection regulation. Comput. Law Secur. Rev. **33**, 171–181 (2017)

Cavoukian, A.: The 7 Foundational Principles. Identity in the Information Society (2010)

Clifton, C., Kantarcioglu, M., Vaidya, J.: Defining Privacy for Data Mining, in National Science Foundation Workshop on Next Generation Data Mining, Baltimore, MD, pp 126–133, November 2002

Dinur, I., Kobbi, N.: Revealing information while preserving privacy. In: Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (2003)

Domingo-Ferrer, J., Montes, F.: Privacy in statistical databases, PSD. In: International Conference on Privacy in Statistical Databases, UNESCO Chair in Data Privacy, International Conference, PSD 2018, Valencia, Spain, 26–28 September 2018, Proceedings (2018)

Domingo-Ferrer, J., Sánchez, D., Blanco-Justicia, A.: The limits of differential privacy (and its misuse in data release and machine learning) (2011)

Drąg, P., Szymura, M.: Technical and legal aspects of database's security in the light of implementation of General data protection regulation. In: CBU International Conference on Innovation in Science and Education (2018)

Dwork, C.: The promise of differential privacy: a tutorial on algorithmic techniques. In: Proceedings - Annual IEEE Symposium on Foundations of Computer Science, FOCS (2011)

[42] Art. 2, point (10) "'data altruism' means the consent by data subjects to process personal data pertaining to them, or permissions of other data holders to allow the use of their non-personal data without seeking a reward, for purposes of general interest, such as scientific research purposes or improving public services", and art. 15 "Register of recognised data altruism organisations. (1) Each competent authority designated pursuant to Article 20 shall keep a register of recognised data altruism organisations. (2) The Commission shall maintain a Union register of recognised data altruism organisations. (3) An entity registered in the register in accordance with Article 16 may refer to itself as a 'data altruism organisation recognised in the Union' in its written and spoken communication."

Elliot, M., Mackey, E., O'Hara, K., Tudor, C.: The anonymization decision-making framework. In: Brussels Privacy Symposium, vol. 1 (2016)
Elliot, M., Domingo Ferrer, J.: The future of statistical disclosure control. Paper published as part of The National Statistician's Quality Review, London, December 2018
Elliot, M., et al.: Functional anonymization: personal data and the data environment. Comput. Law Secur. Rev. **34**(2) (2018)
Finck, M., Pallas, F.: They who must not be identified—distinguishing personal from non-personal data under the GDPR. Int. Data Priv. Law **10**(1) (2020)
Gellert, R.: Understanding the notion of risk in the general data protection regulation. Comput. Law Secur. Rev. **34**(2) (2018)
Graef, I., Gellert, R., Husovec, M.: Towards a holistic regulatory approach for the european data economy: why the illusive notion of non-personal data is counterproductive to data innovation. SSRN Electron. J. (2018)
Hu, R., Stalla-Bourdillon, S., Yang, M., Schiavo, V., Sassone, V.: Bridging policy, regulation and practice? A techno-legal analysis of three types of data in the GDPR (2017)
Hundepool, A., Willenborg, L.: μ- and T-argus: software for statistical disclosure control. In: Third International Seminar on Statistical Confidentiality, Bled (1996)
Jakob, C.E.M., Kohlmayer, F., Meurers, T., Vehreschild, J.J., Prasser, F.: Design and evaluation of a data anonymization pipeline to promote Open Science on COVID-19. Sci. Data **7**, Article no. 435 (2020)
Lane, J., Stodden, V., Bender, S., Nissenbaum, H.: Privacy, Big Data, and the Public Good. Privacy, Big Data, and the Public Good (2014). https://doi.org/10.1017/cbo9781107590205
Leenes, R.: Do you know me? – deconstructing identifiability. Univ. Ott. Law Technol. J. **4**(1&2) (2008)
Li, N., Tiancheng, L., Venkatasubramanian, S.: t-closeness: privacy beyond k-anonymity and l-diversity. In: ICDE (2007)
Le Fevre, K., DeWitt, D.J., Ramakrishnan, R.: Incognito: efficient full-domain k-anonymity. In: SIGMOD Conference (2005)
Machanavajjhala, A., Kifer, D., Kifer, D., Gehrke, J., Gehrke, J., Venkitasubramaniam, M.: L-diversity: privacy beyond k-anonymity. ACM Trans. Knowl. Discov. Data (2007)
Mourby, M., et al.: Are 'pseudonymised' data always personal Data? Implications of the GDPR for administrative data research in the UK. Comput. Law Secur. Rev. **34**(2) (2018)
Ohm, P.: Broken promises of privacy: responding to the surprising failure of anonymization. UCLA Law Rev. **57**(6) (2010)
Palmirani, M., Martoni, M.: Big data, data governance, and new vulnerabilities [big data, governance dei dati e nuove vulnerabilità]. Notizie Di Politeia (2019)
Perera, C., Ranjan, R., Wang, L., Khan, S., Zomaya, A.: Big data privacy in the Internet of Things era. IT Prof. (2015)
Pfitzmann, A., Hansen, M.: A terminology for talking about privacy by data minimization: anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management, Technical University Dresden (2010)
Purtova, N.: The law of everything. broad concept of personal data and future of EU data protection law. Law Innov. Technol. **10**(1) (2018)
Rubinstein, I.S., Hartzog, W.: Anonymization and risk. Wash. Law Rev. **91**(2) (2016)
Samarati P., Sweeney, L.: Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Harv. Data Priv. Lab. (1998)
Samarati, P.: Protecting respondents' identities in microdata release. IEEE Trans. Knowl. Data Eng. (2001)
Sollins, K.: IoT big data security and privacy versus innovation. IEEE Internet Things J. (2019)

Stalla-Bourdillon, S., Knight, A.: Anonymous data v. personal data—a false debate: an EU perspective on anonymization, pseudonymization and personal data. Wis. Int. Law J. **34**(2) (2017)

Stevens, L.: The proposed data protection regulation and its potential impact on social sciences research in the UK. Eur. Data Prot. Law Rev. (2017)

Sweeney, L.: Computational disclosure control: a primer on data privacy protection, Ph.D. thesis, Massachusetts Institute of Technology (2001)

Sweeney, L.: Information explosion. In: Zayatz, L., Doyle, P., Theeuwes, J., Lane, J. (eds.) Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, Urban Institute, Washington, DC (2001)

Wing, J.M.: The data life cycle. Harv. Data Sci. Rev. (2019)

Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., Fu, A.W.-C.: Utility-based anonymization using local recoding. In: KDD (2006)