Bridging Policy, Regulation and Practice?
A techno-legal Analysis of Three Types of Data in the GDPR

Runshan Hu, Sophie Stalla-Bourdillon, Mu Yang, Valeria Schiavo and Vladimiro Sassone

Abstract : The paper aims to determine how the General Data Protection Regulation (GDPR) could be read in harmony with Article 29 Working Party's Opinion on anonymisation techniques. To this end, based on an interdisciplinary methodology, a common terminology to capture the novel elements enshrined in the GDPR is built, and, a series of key concepts (i.e. sanitisation techniques, contextual controls, local linkability, global linkability, domain linkability) followed by a set of definitions for three types of data emerging from the GDPR are introduced. Importantly, two initial assumptions are made: 1) the notion of identifiability (i.e. being identified or identifiable) is used consistently across the GDPR (e.g. Article 4 and Recital 26); 2) the Opinion on Anonymisation Techniques is still good guidance as regards the classification of re-identification risks and the description of sanitisation techniques. It is suggested that even if these two premises seem to lead to an over-restrictive approach, this holds true as long as contextual controls are not combined with sanitisation techniques. Yet, contextual controls have been conceived as complementary to sanitisation techniques by the drafters of the GDPR. The paper concludes that the GDPR is compatible with a risk-based approach when contextual controls are combined with sanitisation techniques.

1. **Introduction**

In recent years, the debate about personal data protection has intensified as a result of an increasing demand for consistent and comprehensive protection of personal data leading to the adoption of new laws in particular in the European Union (EU). The current EU data protection legislation, Data Protection Directive 95/ 46/EC (DPD),[1] is to be replaced by the General Data Protection Regulation (GDPR)[2] from 25 May 2018, which, being a self-executing norm, will be directly applicable in all the Member States in the EU. This legislative reform has generated repeated discussions about its potential impact on business processes and procedures as the GDPR contains a number of new provisions intended to benefit EU data subjects and comprises a strengthened arsenal of sanctions, including administrative fines of up to 4% of total worldwide annual turnover of the preceding financial year, for non-compliant data controllers and processors.

One key question is to what extent the GDPR offers better tools than the DPD to frame or confine data analytics as well as data sharing practices. Addressing this issue requires first of all delineating the scope of data protection law. Second, it necessitates examining key compliance techniques, such as pseudonymisation, of which the raison d'être is to enable data controllers to strike an appropriate balance between two distinct regulatory objectives: personal data protection and data utility maximisation. Not to be misleading, these challenges are not specific to the GDPR and will arise each time law-makers are being tasked with designing a framework aimed at marrying a high degree of personal data protection with some incentives to exploit the potential of data.

---

[1] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data, 1995 O.J. (L 281) 23/11/1995, p. 31- 50 (EU), at Recital 26 [hereinafter DPD].

[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), 2016 O.J. (L 119) 4.5.2016, p. 1–88 (EU), at Recital 26 [hereinafter GDPR].

Within the GDPR, Articles 2 and 4 are starting points in order to demarcate the material scope of EU data protection law. Under Article 4(1), personal data means:

> any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;

Recital 26 further expands upon the notion of identifiability and appears to draw a distinction between personal data and anonymous information, with anonymous information being excluded from the scope of the GDPR. It is true that this key distinction was already present in the DPD. Nonetheless, the GDPR goes further than the DPD in that it indirectly introduces a new category of data as a result of Article 4,[3] i.e. data that has undergone pseudonymisation, which we will name pseudonymised data, to use a shorter expression, although the former is more accurate than the latter for it implies that the state of the data is not the only qualification trigger.[4] Under Article 4(5) pseudonymisation 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 organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person;

While the final text of the GDPR does not seem at first glance to create an ad hoc regime with fewer obligations for data controllers when they deal with pseudonymised data, Recital 29 specifies:

> In order to create incentives to apply pseudonymisation when processing personal data, measures of pseudonymisation should, whilst allowing general analysis, be possible within the same controller when that controller has taken technical and organisational measures necessary to ensure, for the processing concerned, that this

[3] GDPR, supra note 2, at Article 4(5)
[4] Sophie Stalla-Bourdillon and Alison Knight. "Anonymous data v. Personal data–A false debate: An EU perspective on anonymisation, pseudonymisation and personal data." *Wisconsin International Law Journal* (2017): 284-322.

> Regulation is implemented, and that additional information for attributing the personal data to a specific data subject is kept separately.

Furthermore, Article 11 of the GDPR is worth mentioning as it seems to treat with favours a third category of data, which we name Art.11 data for the sake of the argument. Art.11 data under Article 11[5] of the GDPR, is data so that "the [data] controller is able to demonstrate that it is not in a position to identify the data subject."

Examining the GDPR a couple of questions therefore emerges: whether and when pseudonymised data can become anonymised data and whether and when pseudonymised data can be deemed to be Art. 11 data as well.

A number of legal scholars have been investigating the contours of personal data under EU law, and have proposed refined categories, creating on occasion a spectrum of personal data, more or less complex.[6] The classifications take into account the intactness of personal data (including direct and indirect identifiers[7]) and legal controls to categorise data. For instance, with masked direct identifiers and intact indirect identifiers, data is said to become 'protected pseudonymous data' when legal controls are put in place.[8]

---

[5] GDPR, supra note 2, at Article 11. It is true that Article 11 adds that if the data subject "provides additional information enabling his or her identification," Articles 15 to 20 become applicable. As the data subject is described as the one in possession of the additional information (and not the data controller), Art. 11 data and pseudonymised data should not necessarily be equated.

[6] Khaled El Emam, Eloise Gratton, Jules Polonetsky, Luk Arbuckle, "The Seven States of Data: When is Pseudonymous Data Not Personal Information?", accessed March 13, 2017. https://fpf.org/wp-content/uploads/2016/05/states-v19-1.pdf. [hereinafter The Seven States of Data]; Polonetsky, Jules, Omer Tene, and Kelsey Finch. "Shades of Gray: Seeing the Full Spectrum of Practical Data De-Identification." (2016) 56 , 3 Santa Clara Law Review 593; Mike Hintze, "Viewing The GDPR Through A De-Identification Lens: A Tool For Clarification And Compliance", <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2909121> [accessed March 13, 2017]. See also Paul M. Schwartz, Daniel J. Solove. "The PII problem: Privacy and a new concept of personally identifiable information." *NYUL rev.* 86 (2011): 1814; Khaled El Emam, "Heuristics For De-Identifying Health Data", *IEEE Security & Privacy Magazine*, 6/4 (2008), 58-61.

[7] Tore Dalenius,"Finding a needle in a haystack or identifying anonymous census records." *Journal of official statistics* 2, no. 3 (1986): 329.

[8] The Seven States of Data, supra 6, at 6.

We suggest in this paper that these approaches logically rely upon a pre-GDPR understanding of 'pseudonymisation,' which should not be confused with GDPR Article 4 definition and thereby have not necessarily derived the implications of the new legal definitions emerging from the GDPR.

Article 29 Data Protection Working Party (Art. 29 WP) did provide a comprehensive analysis of data anonymisation techniques[9] in the light of the prescriptions of the DPD. For this purpose, Art. 29 WP identified three common risks and tested the robustness of data anonymisation techniques against these risks. However, as aforementioned this was done in 2014 against the background of the DPD and the relationship between these techniques and the data categories defined in the GDPR has not been analysed yet.

The objective of this paper is therefore to derive the implications of the new legal definitions to be found more or less explicitly in the GDPR and determine how the GDPR could be read in harmony with Art. 29 WP's position, in order to inform the work of researchers, practitioners, and ultimately policy and law-makers. To this end, we built a common terminology to capture the novel elements enshrined in the GDPR and thereby introduce a series of key concepts -sanitisation techniques, contextual controls, local linkability, global linkability, domain linkability- followed by a set of definitions for the three types of data emerging from the GDPR developed on the basis of these key concepts. The methodology implemented to create this terminology is interdisciplinary in nature. It combines a systematic analysis of hard law and soft law instruments -the GDPR, the DPD, Court of Justice of the European Union (CJEU) case law, Art. 29 WP opinion- with a review and assessment of key techniques available to data scientists.

[9] Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymisation Techniques (European Comm'n, Working Paper No. 216, 0829/14/EN, 2014) [hereinafter Opinion on Anonymisation Techniques].

We conclude that, assuming the trichotomy of re-identification risks enumerated by Art. 29 WP should still guide the analysis post-GDPR, the GDPR makes the deployment of a risk-based approach possible as long as contextual controls are combined with sanitisation techniques and a relativist approach to data protection law is adopted.

Consequently, the main contributions of the paper are the following:

a) We offer a granular analysis of the three types of risks to be taken into account in order to assess the robustness of sanitisation techniques. The risks include singling out, linkability and inference, with linkability being split into local, global and domain linkability.
b) We propose a classification of data sanitisation techniques and contextual controls in relation to the three categories of data found in the GDPR.
c) We derive criteria for selecting sanitisation techniques and contextual controls, based on the three types of risks in order to assess the feasibility of a risk-based approach.

Importantly, the two premises of the paper are the following: 1) we assume that the notion of identifiability (i.e. being identified or identifiable) is used consistently across the GDPR (e.g. in Article 4 and in Recital 26); 2) we assume that the Opinion on Anonymisation Techniques is still good guidance as regards the distinction drawn between the three types of re-identification risks and the description of sanitisation techniques. Obviously, both of these premises can be criticised as the GDPR has not been litigated yet and the Opinion on Anonymisation Techniques has been appraised critically for several reasons.[10] However, we suggest that even if these two premises seem to lead to an over-restrictive approach, this holds true as long as contextual controls are not combined with sanitisation techniques. Yet, contextual controls such as technical and

[10] See in particular Khaled El Emam, Cecilia Álvarez, "A critical appraisal of the Article 29 Working Party Opinion 05/2014 on data anonymization techniques". *International Data Privacy Law* 2015, 5 (1): 73-87.

organisational measures have been conceived as complementary to sanitisation techniques by the drafters of the GDPR. Contextual controls, including confidentiality obligations, are thus crucial to move towards a workable risk-based approach as well as a relativist approach to data protection law in general.

Structure of the paper. In Section 2 we sketch the new EU data protection legal framework, i.e. the GDPR, give an overview of three risks identified by Art. 29 WP in relation to identification and identifiability, and define the key components of our common terminology. In Section 3, we unfold our risk-based approach for characterising the three types of data emerging from the GDPR and thereby derive an additional set of definitions. The classification of data sanitisation techniques and contextual controls is then realised in Section 4, followed by our conclusions in Section 5.

1. The Three Types of Data

As aforementioned, three types of data seem to emerge from the analysis of the GDPR. We define them in section 2.1 and then conceptualise the three types of risks identified by Art. 29 WP to assess data anonymisation and masking techniques, which we include within the broader category of sanitisation techniques in section 2.2 and distinguish from contextual controls.

1.1 The GDPR Definitions

The definitions presented in this section are derived from the GDPR, including Recital 26 for Anonymised data, Article 4 for Pseudonymised data, and Article 11 for Art.11 data.

- **'Anonymised data'** means data that "does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable."[11]

[11] GDPR, supra note 2, at Recital 26.

- **'Pseudonymised data'** means personal data that have been processed "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 organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person."[12]
- **'Art.11 data'** means data so that the data controller is "not in a position to identify the data subject"[13] given such data.

The notions of 'identified' and 'identifiable' thus appear of paramount importance to distinguish the different types of data and determine whether a category should be considered personal data. An individual is usually considered identified if the data can be linked to a unique real world identity.[14] As per Recital 26, account should be "taken of all the means reasonably likely to be used either by the [data] controller or by another person directly or indirectly."[15] The term "identifiable" refers to the capability to identify an individual, who is not yet identified, but is described in the data in such a way that if research is conducted using additional information or background knowledge she can then be identified. Arguably, following the GDPR, the same 'means test' (of Recital 26) should apply here as well. The foregoing explains why pseudonymised data is still (at least potentially) considered to be personal data. Recital 26 specifies that "[p]ersonal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person."

[12] GDPR, supra note 2, at Article 4(5).
[13] GDPR, supra note 2, at Article 11.
[14] The Seven States of Data, supra 6.
[15] GDPR, supra note 2, at Recital 26.

While the two concepts of pseudonymised data and Art.11 data overlap (so as Art.11 data and anonymised data as it will be explained below), in order to test the extent to which they actually overlap it is necessary to start by conceiving them differently. Besides, Article 11 does not expressly refer to pseudonymisation.

Sticking to the words of GDPR Article 4, we therefore suggest that in order to characterise data as pseudonymised data one has to determine whether individuals are identifiable once the additional information has been isolated and separated from the dataset. Furthermore, to determine whether individuals are identifiable once the additional information has been isolated and separated from the dataset, only the dataset at stake should be considered. This is why, as it will be explained below, the concept of pseudonymised data is intimately linked to that of local linkability.[16]

On the other hand, in order to characterise data as Art.11 data, one has to determine whether a data controller is in a position to identify individuals, i.e. whether individuals are identifiable given the data controller's capabilities, which should require considering all the datasets in the possession of the data controller; but the data controller's capabilities only (therefore to the exclusion of third parties' capabilities). This is the reason why we suggest that the concept of Art.11 data is intimately linked to that of domain linkability.

Consequently, following this logic we argue that to characterise data as pseudonymised data or Art.11 data it is not enough to point to the fact that the individuals are not directly identified within the dataset at stake. As a result, data controllers should not be entitled not to comply with Articles 15 to 20 simply based on the fact that they have decided not to collect direct identifiers for the creation of the dataset at stake.

[16] Sophie Stalla-Bourdillon and Alison Knight. "Anonymous data v. Personal data–A false debate: An EU perspective on anonymisation, pseudonymisation and personal data." *Wisconsin International Law Journal* (2017): 284-322.

1.1.1 Additional information

As hinted above, the concept of 'additional information' is closely related to that of pseudonymised data. Indeed, it can make data subjects identified or identifiable if combined with pseudonymised data. The GDPR requires it to be kept separately and be subject to technical and organisational measures. A typical example of additional information is the encryption key used for encrypting and decrypting data such as attributes: the encrypted data thus becomes pseudonymised data when the key is separated and subject to technical and organisational measures such as access restriction measures.

Two other important concepts related to additional information are that of 'background knowledge' and 'personal knowledge.'[17] In order to analyse re-identification risk properly, it is crucial to draw a distinction between additional information, background knowledge and personal knowledge.

As per GDPR Article 4, Additional information, is the information that can be kept separately from the dataset by technical and organisational measures, such as encryption key, hash function etc.

We distinguish additional information from background knowledge and personal knowledge. Background knowledge, is understood as different in kind from additional information as it corresponds to knowledge that is publicly accessible to an average individual who is deemed reasonably competent to access it, therefore most likely including the data controller himself. It comprises information accessible through the Web such as news websites or information found in public profiles of individuals or traditional newspapers. While this kind of knowledge can potentially have a high impact on re-identification risks, it cannot be physically separated from a

[17] Information Commissioner's Office, *Anonymisation: Managing Data Protection Risk Code Of Practice*, 2012.

dataset. Therefore, we exclude it from additional information. However, and this is important, we take it into account when we analyse the three types of data by acknowledging that the potential existence of background knowledge makes it necessary to include singling out as a relevant risk for pseudonymised data within the meaning of the GDPR because as a result of a pseudonymisation process, the data shall not be attributable to an identifiable data subject as well. The same is true for Art. 11 data.[18]

Personal knowledge, is assessed through the means of a subjective test (as opposed to background knowledge, which is assessed through the means of an objective test) and varies from one person to another.[19] It comprises information that is not publicly accessible to an average individual who is deemed reasonably competent to access it, but only to certain individuals because of their special characteristics. For example, a motivated intruder A has the knowledge that B is currently in hospital, as she is B's neighbour and she saw that B was picked up by an ambulance. When combined with anonymised data, this kind of subjective personal knowledge could obviously result in re-identification. However, for the purposes of this paper we assume that the likelihood that a motivated intruder has relevant personal knowledge is negligible, which partly depends upon his/her willingness to acquire this relevant personal knowledge and his/her estimation of the value of the data at stake and thereby the degree of data sensitivity. We recognise, however, that further sophistication would be needed for scenarios in which the likelihood that a motivated intruder has relevant personal knowledge is high. In particular, this would mean considering with care the equivalence of sanitisation techniques and contextual controls. With this

[18] It might be that a less restrictive approach would be preferable but the purpose of this paper is to show that the restrictiveness of the approach can ultimately be mitigated with contextual controls.
[19] Information Commissioner's Office, *Anonymisation: Managing Data Protection Risk Code Of Practice*, 2012.

said, we note that Art. 29 WP wrote in 2007 that " a mere hypothetical possibility to single out the individual is not enough to consider the person as "identifiable"."[20]

1.1.2 Direct and indirect identifiers

As described in the ISO/TS document, direct identifier is "data that can be used to identify a person without additional information or with cross-linking through other information that is in the public domain."[21] Direct identifiers contain explicitly identifying information, such as names and social security numbers that are uniquely linked to a data subject. In contrast, sets of attributes which can be combined together to uniquely identify a data subject, are called indirect identifiers. They include age, gender, zip code, date of birth and other basic demographic information. No single indirect identifier can identify an individual by its own; however, the re-identification risks appear when combining indirect identifiers together, as well as, as aforementioned, when combining records with additional information or with background knowledge. Notably, the list of direct and indirectly identifiers can only be derived contextually.

1.1.3 Data sanitisation techniques

Data sanitisation techniques process data in a form that aims to prevent re-identification of data subjects. Randomisation and generalisation are considered as two main families of sanitisation techniques.[22] There is a wide range of techniques including masking techniques, noise addition, permutation, k-anonymity, l-diversity and differential privacy, etc. Noise addition refers to general techniques that make data less accurate by adding noise usually bounded by a range, e.g., [-10, 10]. We differentiate it from differential privacy as the latter offers more rigorous guarantee.

[20] Article 29 Data Protection Working Party, Opinion 04/2007 on the concept of personal data (European Comm'n, Working Paper No. 136, 01248/07/EN), p. 15.
[21] International Organization for Standardization, *ISO/TS 25237:2008 Health Informatics – Pseudonymization*, 2008 <https://www.iso.org/standard/42807.html> [accessed 13 March 2017].
[22] Opinion on Anonymisation Techniques, supra note 9, at 12.

Masking or removal techniques are applied to direct identifiers to make sure the data subjects are not identified anymore and then additional techniques (including masking techniques) are then used to further process indirect identifiers. It is true that k-anonymity, l-diversity, and differential privacy are more commonly described as privacy models rather than techniques as such. However, as we built upon the Opinion on Anonymisation Techniques we use a similar terminology to simplify the arguments.

1.1.4 Contextual controls

Contextual controls comprise three sets of controls. First, legal and organisational controls such as obligations between parties and/or internal policies adopted within one single entity (one party) aimed at directly reducing re-identification risks, e.g. obligation not to re-identify or not to link. Second, security measures (including legal, organisational and technical controls) such as data access monitoring and restriction measures, auditing requirements as well as additional security measures, such as the monitoring of queries, all of them aimed at ensuring the de facto enforcement of the first set of controls. Third, legal, organisational and technical controls relating to the sharing of datasets aimed at ensuring that the first set of legal controls are transferred to recipients of datasets. They include obligations to share the datasets with the same set of obligations or an obligation not to share the datasets, as well as technical measures such as encryption to make sure confidentiality of the data is maintained during the transfer of the datasets.

These measures are used to balance the strength of data sanitisation techniques with the degree of data utility. In this sense, they are complementary to data sanitisation techniques. On one hand, they reduce residual risks, which remain after implementing data sanitisation techniques; on the other hand, they make it possible to preserve data utility while protecting the personal data of data subjects.

In practice, the selection of contextual controls depends on specific data sharing scenarios.

2.2 Re-Identification Risks

The re-identification risks relate to ways attackers can identify data subjects within datasets. Art. 29 WP's Opinion on Anonymisation Techniques[23] describes three common risks and, examines the robustness of data sanitisation techniques against those risks.[24] Underlying this risk classification is the premise that the means test is a tool to "assess whether the anonymisation process is sufficiently robust."[25]

- **'Singling out'**, which is the "possibility to isolate some or all records which identify an individual in the dataset."[26]
- **'Linkability'**, which is the "ability to link at least two records concerning the same data subject or a group of data subjects (either in the same database or in two different databases)."[27]
- **'Inference'**, which is the "possibility to deduce, with significant probability, the value of an attribute from the values of other attributes."[28]

In cases in which there is background knowledge, singling out makes an individual identifiable. The connection between identifiability and linkability or inference is less straightforward. Adopting a restrictive approach one could try to argue that if background knowledge exists so that it is known that an individual belongs to a grouping in a dataset, the

[23] Opinion on Anonymisation Techniques, supra note 9, at 11-12.
[24] As hinted above, it maybe that this classification needs to be re-thought as for example it does not distinguish between attribute disclosure and identity disclosure. This not, however, the purpose of this paper.
[25] Opinion on Anonymisation Techniques, supra note 9, at 8.
[26] Opinion on Anonymisation Techniques, supra note 9, at 11.
[27] Opinion on Anonymisation Techniques, supra note 9, at 11.
[28] Opinion on Anonymisation Techniques, supra note 9, at 12.

inferred attribute(s) combined with background knowledge could lead to identification or at the very least disclosure of (potentially sensitive) information relating to an individual.

Art. 29 WP categorised data sanitisation techniques into 'randomisation', 'generalisation' and 'masking direct identifiers'[29], where randomisation and generalisation are viewed as methods of anonymisation but masking direct identifiers or pseudonymisation (to use the words of Art. 29 WP) as a security measure. It should be clear from now that the GDPR definition of pseudonymisation is more restrictive than merely masking direct identifiers. Masking direct identifiers is conceived as a security measure by Art. 29 WP because it does not mitigate the three risks aforementioned; or rather, it simply removes/masks the direct identifiers of data subjects.

'Noise addition', 'permutation' and 'differential privacy' are included within the randomisation group as they alter the veracity of data. More specifically, noise addition and permutation can reduce linkability and inference risks, but fail to prevent the singling out risk. Differential privacy is able to prevent all the risks up to a maximum number of queries or until the predefined privacy budget is exhausted but queries must be monitored and tracked when multiple queries are allowed on a single dataset. As regards the generalisation category, 'K-anonymity'[30] is considered robust against singling out, but linkability and inference risks are still present. 'L-diversity'[31] is stronger than K-anonymity provided it first meets the minimum criterion of k-anonymity, as it prevents both the singling out and inference risks.

Although Art. 29 WP has provided insights for the selection of appropriate data sanitisation techniques, which are relevant in the context of personal data sharing, these techniques ought to

[29] Opinion on Anonymisation Techniques, supra note 9, at 12.
[30] Latanya Sweeney, "K-Anonymity: A Model For Protecting Privacy", *International Journal Of Uncertainty, Fuzziness And Knowledge-Based Systems*, 10/05 (2002), 557-570.
[31] Ashwin Machanavajjhala and others, "L-Diversity", *ACM Transactions On Knowledge Discovery From Data*, 1/1 (2007).

be examined in the light of the GDPR. To be clear, the purpose of this paper is not to question the conceptualisation of re-identification risks undertaken by Art. 29 WP, but to deduce its implications when interpreting the GDPR in context.

2. A Risk-based Analysis of the Three Types of Data

In this section, we refine the concept of linkability and further specify the definitions of the three categories of data emerging from the GDPR using a risk-based approach.

2.1 Local, Global and Domain Linkability

Analysing in a more granular fashion the linkability risk defined by Art. 29 WP, it is possible to draw a distinction between three scenarios. The first scenario focuses on a single dataset, which contains multiple records about the same data subject. An attacker identifies the data subject by linking these records using some additional information. In the second scenario, the records of a data subject are included in more than one datasets, but these datasets are held within one entity. An attacker links the records of a data subject if she can access all the datasets inside the entity, e.g., insider threat.[32] The third scenario also involves more than one datasets, but these datasets are not necessarily held within one entity. Based on these three scenarios, we distinguish between three types of linkability risks:

- **'Local Linkability',** which is the ability to link records that correspond to the same data subject within the same dataset.
- **'Domain linkability',** which is the ability to link records that correspond to the same data subject in two or more datasets that are in the possession of the data controller.

[32] Theoharidou Marianthi and others, "The Insider Threat To Information Systems And The Effectiveness Of ISO17799", *Computers & Security*, 24/6 (2005), 472-484.

- **'Global Linkability',** which is the ability to link records that correspond to the same data subject in any two or more datasets.

Based on this granular analysis of the linkability risk and assuming the concept of identifiability is used consistently across the GDPR, we suggest one way to derive the main characteristics of anonymised, pseudonymised and Art. 11 data within the meaning of the GDPR.

2.2 Anonymised Data

Anonymised data, according to the GDPR definition, is a state of data for which data subjects are not identified nor identifiable anymore, taking into account all the means reasonably likely to be used by the data controller as well as third parties. While strictly speaking the legal test to be found in Recital 26 of the GDPR does not mention all of the three risks aforementioned (i.e. singling out, linkability and inference), we assume for the purposes of this paper that for anonymised data to be characterised, singling out, local linkability, domain linkability, global linkability and inference should be taken into account. As aforementioned, whether the three re-identification risks should be re-conceptualised is a moot point at this stage. Suffice it note that not all singling out, linkability and inference practices lead to identifiability and identification. A case-by-case approach is therefore needed.

2.3 Pseudonymised Data

Pseudonymised data, being the outcome of the pseudonymisation process defined by the GDPR in its Article 4, is a state of data for which data subjects can no longer be identified or identifiable when examining the dataset at stake (and only the dataset at stake). Nevertheless, the foregoing holds true on the condition that data controllers separate the additional information and

put in place "technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person."

As a result, it appears that pseudonymisation within the meaning of the GDPR is not tantamount to masking direct identifiers. In addition, although a number of studies stress the importance of legal controls,[33] there are different routes to pseudonymised data depending upon the robustness of the sanitisation technique implemented, as it is explained below.

One important element of the GDPR definition of pseudonymisation is the concept of additional information, which can identify data subjects if combined with the dataset. The definition specifies that such additional information is kept separately and safeguarded, so that the risks relating to the additional information can be excluded. This seems to suggest that in this context the notion of identifiability should only relate to the dataset at stake. Based on this analysis, we define pseudonymised data as a data state for which the risks of singling out, local linkability and inference should be mitigated. At this stage, the domain and global linkability risks are not relevant and the data controller could for example be in possession of other types of datasets.

In order to mitigate the singling out, local linkability and inference risks at the same time, data sanitisation techniques must be selected and implemented on the dataset. As aforementioned, Art. 29 WP has examined several sanitisation techniques in relation to re-identification risks.[34] We build on the upshot of the Opinion on Anonymisation Techniques, and find that K-anonymity, L-diversity and other stronger techniques can prevent these risks, but masking direct identifiers, noise

[33] See e.g. The Seven States of Data, supra 6; Polonetsky, Jules, Omer Tene, and Kelsey Finch. "Shades of Gray: Seeing the Full Spectrum of Practical Data De-Identification." (2016) 56, 3 Santa Clara Law Review 593.
[34] Opinion on Anonymisation Techniques, supra note 9, at 13-21.

addition, permutation alone are insufficient to reasonably mitigate the singling out, local linkability and inference risks.

The example below illustrates the mitigation of these three risks using K-anonymity.

Example. Table 1 shows a sanitised dataset with k-anonymity guarantee (k=4) released by hospital A in May. Suppose an attacker obtains relevant background knowledge from a news website that a famous actor Bob was recently sent to hospital A and that by checking the time it can be deduced that Bob is in the dataset at stake. Suppose as well that the attacker has no access to additional information (e.g. the raw dataset). Since each group of this dataset has at least 4 records sharing the same non-sensitive attribute values, the attacker cannot distinguish his target Bob from other records. This prevents the risks of singling out and local linkability. Moreover, the attacker is not able to infer the sensitive attribute of Bob because she is not sure to which group Bob belongs. Therefore, this dataset is pseudonymised within the meaning of the GDPR.

Table 1 An example of Pseudonymised data using k-anonymity (k=4)

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Diagnosis
1	250**	<30	*	Cancer
2	250**	<30	*	Viral Infection
3	250**	<30	*	AIDS
4	250**	<30	*	Viral Infection
5	250**	3*	*	Cancer
6	250**	3*	*	Flu
7	250**	3*	*	Cancer
8	250**	3*	*	Flu

2.4 Art. 11 Data

Art. 11 data, by definition, focuses on the ability of a data controller to identify data subjects to the exclusion of third parties. More specifically, the data controller should be able to

demonstrate that she is "not in a position to identify the data subject."[35] First, this implies that direct identifiers (e.g. names, social security number etc.) have been removed or have never been collected. In other words, Art. 11 data is either sanitised by a certain process or not. Second, "not being in a position to identify the data subject" should also imply that the combination of indirect identifiers does not lead to identification. There exist also situations where data controller only collect indirect identifiers but a very rich of list of indirect identifiers for which arguably, and this is crucial, no accessible relevant background knowledge exists and the data controller is not in a possession of other datasets which could be linked to the first one, e.g. dynamic IP addresses, browsed websites and search terms, transactions… in order to create profiles and ultimately make decisions about individuals. We suggest that while an approach purely based on a re-identification risks approach would lead to exempting data controllers from Articles 15 to 20 in these situations, this would not necessarily be consistent with the spirit of the GDPR, which aims to strengthen the protection of data subjects in cases of profiling. As a result, in order to determine whether data is personal data and the full data protection regime applies two scenarios must be taken into account: 1) whether re-identification risks have been appropriately mitigated and 2) whether profiling and decisions about individuals are made.

Importantly, Art. 11 definition requires that to determine whether the data is Art. 11 data, all the means of the data controller should be considered to the exclusion of third parties' means. As a result, Art. 11 data can be interpreted as a state of data for which there are no risks of singling out, domain linkability and inference. The protection applied to Art. 11 data is therefore stronger than the protection applied to pseudonymised data because the former requires mitigating the domain linkability rather than local linkability risk. This does not mean that pseudonymised data

[35] GDPR, supra note 2, at Article 11.

cannot be transformed into Art. 11 data. The example below illustrates the difference between Art. 11 and pseudonymised data.

Example. Suppose two hospitals H_1 and H_2 located in a same city publish patient data frequently, e.g., weekly. Table 2(a) is the dataset sanitised and published by H_1 using k-anonymity (k=4). The dataset achieves the state of pseudonymised data as no record in the table can be attributed to a specific data subject without using additional information. Furthermore, H_1 claims that it is not able to identify any data subject using any other information within the domain/access of H_1. This other information could be the datasets previously published by H_1 and H_2. One week later, H_2 publishes its own patient dataset. It sanitises the data using k-anonymity (k=6) and achieves the state of pseudonymised data, as shown in Table 2(b). Now H_2 wants to determine whether the dataset (Table 2(b)) is also Art. 11 data. H_2 is in possession of other information (different from the concept of additional information) comprising Table 2(a), and background knowledge deriving from a news website (which has been read by many people in the city) saying that a 28-year-old celebrity living in zip code 25013 has been sent to both H_1 and H_2 to seek a cure for his illness. H_2 thus goes through the medical records of each patient. With the other information, H_2 knows that the celebrity must be one of the four records in Table 2(a) and one of the six records in Table 2(b). H_2 is therefore able to identify the celebrity by combining Table 2(a) and Table 2(b), because only one patient was diagnosed with the disease that appears in both tables, i.e., cancer. As a result, H_2 can be sure that the celebrity matches the first record of both tables, and the celebrity has cancer. Therefore, Table 2(b) comprises pseudonymised data but not necessarily Art. 11 data.

Table 2(a) 4-anonymous patient data from H_1

	Non-Sensitive			Sensitive
	Zip code	Age	B_city	Diagnosis
1	250**	<30	*	Cancer
2	250**	<30	*	Viral Infection
3	250**	<30	*	AIDS
4	250**	<30	*	Viral Infection
5	250**	3*	*	AIDS
6	250**	3*	*	Heart Disease
7	250**	3*	*	Heart Disease
8	250**	3*	*	Viral Infection
9	250**	≥40	*	Cancer
10	250**	≥40	*	Cancer
11	250**	≥40	*	Flu
12	250**	≥40	*	Flu

Table 2(b) 6-anonymous patient data from H_2

	Non-Sensitive			Sensitive
	Zip code	Age	B_city	Diagnosis
1	250**	<35	*	Cancer
2	250**	<35	*	Tuberculosis
3	250**	<35	*	Heart Disease
4	250**	<35	*	Heart Disease
5	250**	<35	*	Flu
6	250**	<35	*	Flu
7	250**	≥35	*	Heart Disease
8	250**	≥35	*	Viral Infection
9	250**	≥35	*	Flu
10	250**	≥35	*	Flu
11	250**	≥35	*	Flu
12	250**	≥35	*	Flu

We summarise the three types of data based on the risks aforementioned in the following table.

Table 3 Risk-based interpretation for three types of data

	Singling out	Local linkability	domain linkability	Global linkability	Inference
Anonymised data	No	No	No	No	No
Art. 11 data	No	No	No	N/A	No
Pseudonymised data	No	No	N/A	N/A	No

3. Data Sanitisation Techniques and Contextual Controls

We now examine the robustness of data sanitisation techniques against the five types of re-identification risks. Taking into account data sharing contexts, we present a hybrid assessment comprising both contextual controls and data sanitisation techniques.

3.1 Effectiveness of data sanitisation techniques

We build upon the table of data sanitisation techniques presented by Art. 29 WP[36] by splitting the linkability risk into local and global linkability. At this stage, domain linkability is not explicitly shown in the table as it is included in global linkability. The table below summarises the results.

Table 4 Robustness of data sanitisation techniques

	Is singling out still a risk?	Is local linkability still a risk?	Is domain/global linkability still a risk?	Is inference still a risk?
Masking direct identifiers	Yes	Yes	Yes	Yes
Noise Addition	Yes	May not	May not	May not
Permutation	Yes	Yes	Yes	May not
Masking indirect identifiers	Yes	Yes	Yes	May not
K-anonymity	No	No	Yes	Yes
L-diversity	No	No	Yes	May not
Differential privacy	May not	May not	May not	May not

Note that domain linkability is in the same column as global linkability, because for both situations external datasets need to be taken into account and the listed data sanitisation techniques are not able to distinguish between different types of domains. While one should revert to explanations provided by Art. 29 WP[37] for the analysis of the singling out and inference risks, we then discuss the robustness of sanitisation techniques in relation to local, domain and global linkability risks.

Masking direct identifiers. Applying the techniques, such as encryption, hashing and tokenisation on direct identifiers, can reduce linkability between a record and the original identity of a data subject (e.g., name). However, it is still possible to single out data subjects' records with

[36] Opinion on Anonymisation Techniques, supra note 9, at 24.
[37] Opinion on Anonymisation Techniques, supra note 9, at 13-21.

the pseudonymised attributes. If the same pseudonymised attribute is used for the same data subject, then records in one or more datasets can be linked together. If different pseudonymised attributes are used for the same data subject and there is at least one common attribute between records, it is still possible to link records using other attributes. Therefore, the local, domain and global linkability risks exist in both situations.

Noise Addition. This technique adds noise to attributes, making the values of such attributes inaccurate or less precise. However, this technique cannot mitigate local, domain and global linkability risks. Indeed, this technique only reduces the reliability of linking records to data subjects as the values of attributes are more ambiguous. Records may still be linked using wrong attribute values.

Permutation. Permutation is a technique that consists in shuffling values of attributes within a dataset. More specifically, it swaps values of attributes among different records. It can be considered as a special type of noise addition[38] though it retains the range and distribution of the values. Therefore, it is still vulnerable to the local, domain and global linkability risks based on the shuffled values of attributes, although such linking may be inaccurate as an attribute value may be attached to a different subject.

K-anonymity. As the main technique of the generalisation family, K-anonymity is applied to prevent singling out. They group a data subject with at least k-1 other individuals who share a same set of attribute values.[39] These techniques are able to prevent local linkability, because the probability of linking two records to the same data subject is no more than 1/k. However, they are

[38] Opinion on Anonymisation Techniques, supra note 9, at 13.
[39] Latanya Sweeney, "K-Anonymity: A Model For Protecting Privacy", *International Journal Of Uncertainty, Fuzziness And Knowledge-Based Systems*, 10/05 (2002), 557-570.

not able to mitigate the domain and global linkability risks. As shown in our example of the two hospitals, records relating to the celebrity can be linked together via an intersection attack.[40]

L-diversity. Compared with K-anonymity, the significant improvement of L-diversity is that it ensures the sensitive attribute in each equivalence class has at least L different values.[41] Thus, it prevents the risk of inference to the probability of no more than 1/L. However, like K-anonymity, it cannot prevent domain and global linkability as shown in our example of two hospitals because it is still possible to link records together if they have the same sensitive attribute values.

Differential privacy. Differential privacy is one of the randomisation techniques that can ensure protection in a mathematical way by adding a certain amount of random noise to the outcome of queries.[42] Differential privacy means that it is not possible to determine whether a data subject is included in a dataset given the query outcome. In the situation where multiple queries on one or more datasets are allowed, the queries must however be tracked and the noise should be tuned accordingly to ensure attackers cannot infer more information based on the outcomes of multiple queries. Therefore, "May not" is assigned for the risks depending on whether queries are tracked.

Masking indirect identifiers. As described before, encryption, hashing and tokenisation are the techniques for masking direct identifiers. They can also be implemented on indirect

[40] Srivatsava Ranjit Ganta, Shiva Prasad Kasiviswanathan and Adam Smith, "Composition Attacks And Auxiliary Information In Data Privacy", *Proceeding Of The 14Th ACM SIGKDD International Conference On Knowledge Discovery And Data Mining - KDD 08*, 2008.
[41] Ashwin Machanavajjhala and others, "L-Diversity", *ACM Transactions On Knowledge Discovery From Data*, 1/1 (2007), 3-es.
[42] Cynthia Dwork, "Differential Privacy: A Survey Of Results", in *In International Conference On Theory And Applications Of Models Of Computation* (Berlin Heidelberg, 2008), 1-19.

identifiers. We observe that these techniques are not able to mitigate the risks of local, domain and global linkability. Taking a dataset with three quasi-identifiers - gender, address and date of birth, for example, a hash function encrypts the combination of the three quasi-identifiers. If there are two records in the dataset (or different datasets) corresponding to a same data subject, then they will have the same hashed values for these three attributes.

We now combine our risk-based interpretation of three types of data (Table 3) with the foregoing analysis of the robustness of data sanitisation techniques (Table 4), in order to classify the output of different techniques into three types of data.

Table 5 The results of data sanitisation techniques

Techniques	Pseudonymised data	Art. 11 data	Anonymised data
Masking direct identifiers	Not	Not	Not
Noise Addition	Not	Not	Not
Permutation	Not	Not	Not
Masking indirect identifiers	Not	Not	Not
K-anonymity	Not	Not	Not
L-diversity	Yes	Not	Not
Differential Privacy	Maybe	Maybe	Maybe

As the first four techniques are not able to mitigate the risk of singling out, the outcome of these four techniques cannot be pseudonymised data, Art. 11 data, or anonymised data. For K-anonymity, it cannot produce any of these three data types because it only mitigates singling out and local linkability to the exclusion of inference when additional information is isolated and safeguarded. Notably, background knowledge is taken into account. Data after implementing L-diversity is pseudonymised data because it can mitigate singling out, local linkability, and inference, but not domain linkability or global linkability. As for Art. 11 data, L-diversity does not mitigate against the fact that data controllers have within their domain other datasets, which can

be used to link records together. Hence, "Not" is assigned. Differential privacy can guarantee Art. 11 data, pseudonymised data or anonymised data if only single query on one dataset is allowed or multiple queries are tracked.

So far, we have classified data sanitisation techniques with respect to the three types of data. It is worth mentioning that data sanitisation techniques are often combined in practice. Table 5 derives the sanitisation outcome in situations where two or more techniques are implemented. For example, (K, L) - anonymity[43] combining K-anonymity and L-diversity, ensures that each equivalent class has at least K records, and their sensitive attributes have at least L different values. (K, L) - anonymity guarantees that there are no risks of singling out, local linkability and inference.

3.2 Improving data utility with contextual controls

Maintaining an appropriate balance between data utility and data protection is not an easy task for data controllers. As discussed in Section 4.1, K-anonymity, L-diversity and differential privacy are the sole potential techniques that can make data pseudonymised, Art. 11 or anonymised data. However, these techniques could introduce undesired distortion on data, making data less useful for data analysts. Contextual controls are thus crucial to complement data sanitisation techniques and reduce risks.[44] Obviously, the strength of the contextual control to add should depend upon the type of data sharing scenarios at hand.

[43] Ji-Won Byun and others, "Privacy-Preserving Incremental Data Dissemination", *Journal Of Computer Security*, 17/1 (2009), 43-68.

[44] Leibniz Institute for Educational Trajectories (LIfBi), *Star Ng Cohort 6: Adults (SC6) SUF Version 7.0.0 Anonymiza On Procedures Tobias Koberg*, 2009 <https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC6/7-0-0/SC6_7-0-0_Anonymization.pdf> [accessed 13 March 2017].

In order to take into account the variety of data sharing scenarios, we distinguish between two types of contextual legal controls: **'inter-party'** and **'internal controls'.** The former category comprises obligations between parties (i.e. data collector/data releaser and data recipient), and the latter comprises internal policies adopted within one entity, i.e. one party. As shown in Table 6, the top rows of controls are meant to directly address the re-identification risks. The middle rows list the controls used to ensure that the first set of controls are actually implemented. More specifically, security measures are measures that relate to location of storage, access to data, training of staff and enforcement of internal policies. Additional security measures are associated with differential privacy only and are required to guarantee differential privacy mitigates all the risks. The third set of controls is essential when data are shared in order to make sure recipients of datasets put in place the necessary controls to maintain the dataset within its initial category: depending upon the sensitivity of the data they take the form of obligations/policies not to share the data or an obligation to share the data alike, i.e. with the same controls. Technical measures, such as encryption, can complement these obligations to make sure confidentiality of the data is maintained during the transfer of the dataset to the recipients.

Table 6 Inter-party (obligations) and Internal (policies) controls

1. Mitigating risks directly	*Singling out risk* • Obligation/Policy to isolate info to de-mask direct identifiers with security measures in relation to location of storage, access to formula, training of staff and enforcement of rules • Obligation/Policy not to identify from indirect identifiers
	Local linkability risk • Obligation/Policy not to link records in the same dataset
	domain linkability risk Obligation/Policy not to link with other datasets within the same domain
	Global linkability risk • Obligation/Policy not to link with other datasets
	Inference risk

	Obligation/Policy not to infer attributes from existing attributes
2. Enforcing the mitigation	*Security measures* • Obligation/Policy to implement security measures in relation to location of storage, access to dataset, auditing, training of staff and enforcement of internal policy rules
	Additional security measures • Obligation/Policy to monitor queries and query outcome after applying differential privacy
3. Transferring controls	• Obligation/Policy not to re-share or to re-share with the same set of obligations • Obligation to share data in an encrypted state, e.g., through an encrypted communication channel

It is now time to combine data sanitisation techniques and contextual controls to determine when and how it is possible to maintain data utility. This is the objective of Tables 7 and 8.

Two types of actors are distinguished to take into account the implications of data sharing scenarios: *data collectors,* who collect original data and transform the data in certain data types before sharing the data; and *data recipients*, who receive processed data and may have to implement controls in order to ensure the data remain within the desired data category. Table 7 only concerns data collectors. This is why no inter-party controls are considered.

Table 7 Sanitisation options when data are in the hands of data collectors

Desired data type	**Sanitisation options**
Pseudonymised data	• Masking direct identifiers + Policies on singling out, local linkability and inference risks + Security measures • K-anonymity + Policy on inference risk + Security measures • L-diversity + Security measures
Art. 11 data	• Masking direct identifiers/Collecting only indirect identifiers + Policies on singling out, domain linkability risks + Security measures • K-anonymity + Policies on inference and domain linkability risks + Security measures • L-diversity + Policy on domain linkability risk + Security measures

Anonymised data	• Masking direct identifiers + Policies on singling out, local, global linkability and inference risks + Security measures • K-anonymity + Policies on inference and global linkability risks + Security measures • L-diversity + Policies on global linkability risk + Security measures • Differential privacy + Security measures + Additional security measures

In the first row of the table, data fall into the category of pseudonymised data when the singling out, local linkability and inference risks have been mitigated. When implementing a weak sanitisation technique only, i.e. masking direct identifiers, those risks still persist as explained above and contextual controls are therefore needed. Stronger data sanitisation techniques, such as K-anonymity and L-diversity, mitigate more risks, which explains why fewer and/or weaker contextual controls are needed. For instance, when L-diversity is implemented, only security measures are required for achieving pseudonymised data.

In the end the selection of data sanitisation techniques and contextual controls should depend on the type of data sharing scenario pursued (closed or open) given both the sensitivity and the utility of the data.

Data in the second category, i.e. Art. 11 data, implies that the data controller is able to demonstrate that she is not in a position to identify data subjects. The listed options ensure that there are no singling out, domain linkability and inference risks.

Data in the final category is anonymised data, which require the strongest protection, i.e. that no singling out, local and global linkability and inference risks exist. Differential privacy is one of the options, and only security measures are required when differential privacy is implemented.

Table 8 concerns data recipients. As for data recipients who receive processed data, they should take into account (i) the data sanitisation techniques that have been implemented on the received data, and (ii) the obligations imposed by data releasers.

Table 8 Sanitisation options when data are in the hands of data recipients

Desired data type	**Sanitisation techniques implemented on received data**	**Obligations imposed upon data recipients**	**Sanitisation options**
Pseudonymised data	Masking direct identifiers	Obligations on singling out, local linkability and inference risks + obligation on implementing security measures	• Policies on singling out, local linkability and inference risks + Security measures • K-anonymity + Policy on inference risk + Security measures • L-diversity + Security measures
	K-anonymity	Obligation on inference risk + obligation on implementing security measures	• Security measures • L-diversity + Security measures
	L-diversity	Obligation on implementing security measures	• Security measures
Art. 11 data	Masking direct identifiers	Obligations on singling out, inference, local and domain linkability risks + obligation on implementing security measures	• Policies on singling out, inference, local and domain linkability risks + Security measures • K-anonymity + Policies on inference, domain linkability risks + Security measures • L-diversity + Policy on domain linkability risk + Security measures
	K-anonymity	Obligations on inference and domain linkability risks + obligation on	• Policies on inference and domain linkability risks + Security measures

		implementing security measures	• L-diversity + Policy on domain linkability risk + Security measures
	L-diversity	Obligation on domain linkability risk + obligation on implementing security measures	• Policy on domain linkability risk + Security measures
Anonymised data	Masking direct identifiers	Obligations on singling out, local, global linkability and inference risks + obligation on implementing security measures	• Policies on singling out, local, global linkability and inference risks + Security measures • K-anonymity + Policies on inference and global linkability risks + Security measures • L-diversity + Policy on global linkability risk + Security measures • Differential privacy + Security measures + Additional security measures
	K-anonymity	Obligations on inference and global linkability risks + obligation on implementing security measures	• Policies on global linkability and inference risks + Security measures • L-diversity + Policy on global linkability risk + Security measures • Differential privacy + Security measures + Additional security measures
	L-diversity	Obligation on global linkability risk + obligation on implementing security measures	• Policy on global linkability risk + Security measures • Differential privacy + Security measures + Additional security measures
	Differential privacy	Obligation on implementing security measures	• Security measures + Additional security measures

Table 8 provides a number of sanitisation options that data recipients can select to meet their data protection and utility requirements. We take pseudonymised data as an example. Suppose a data recipient receives data that were processed with K-anonymity techniques and she aims to keep the data in a pseudonymised state. The data recipient has thus two options. Either she does not change the data and simply adopt policies and security measures; or she further processes the data with L-diversity, and adopt different types of policies as well as security measures.

Another consideration is worth mentioning. If the data collector keeps the original raw dataset, the original raw dataset should be conceived as falling within the category of additional information for the purposes of characterising personal data and within the category of the data controller's domain for the purposes of characterising Art. 11 data. As regards anonymised data, Art. 29 WP seems to suggest that as long as the raw dataset is not destroyed the sanitised dataset cannot be characterised as anonymised data.[45] Applying a risk-based approach of the type developed in this paper would lead to the opposite result. This said, and this is essential, this would not mean that the data controller transforming and releasing the raw dataset into anonymised data would not be subject to any duty anymore. It would actually make sense to impose upon the data controller a duty to make sure recipients of the dataset put in place the necessary contextual controls. This duty could be performed by imposing upon recipients an obligation not to share the dataset or to share the dataset alike, depending upon data sensitiveness and data utility requirements. Ultimately, the data controller would also be responsible for choosing the appropriate mix of sanitisation techniques and contextual controls as the anonymisation process as such is still a processing activity governed by the GDPR. Data controllers could thus be required to monitor best practices in the field even after the release of the anonymised data.

[45] Opinion on Anonymisation Techniques, supra note 9, at 10.

Finally it should be added that the foregoing analysis implies a relativist approach to data protection law, which would require determining the status of a dataset on a case-by-case basis and thereby for each specific data sharing scenario.

3.3 Improving data utility with dynamic sanitisation techniques and contextual controls

Re-identification risks are not static and evolve over time. This should mean that data controllers should regularly assess these risks and take appropriate measures when their increase is significant.

Notably, adapting sanitisation techniques and contextual controls over time can help reduce re-identification risks. At least one dynamic sanitisation technique is worth mentioning here: changing pseudonyms over time for each use or each type of use as a way to mitigate linkability.[46] Besides, techniques like k-anonymity and l-diversity can also be conceived as dynamic techniques as deploying k or l on the same dataset for new recipients can provide stronger protection when the data controller observes that re-identification risks increase.

At the same time, data recipients should be aware of the limits imposed upon the use of the data, even if the data is characterised as anonymised. This is a logical counterpart to any risk-based approach and necessarily implies that data controllers and data recipients are in continuous direct contact, at least when differential privacy is not opted for. Indeed, contextual controls put in place for mitigating risks directly (in order to preserve data utility) could be coupled with confidentiality obligations and/or confidentiality policy, be it relative (i.e. formulated as an obligation to share alike) or absolute (i.e. formulated as a prohibition to share). Importantly, taking confidentiality

[46] Mike Hintze and Gary LaFever, "Meeting Upcoming GDPR Requirements While Maximizing The Full Value Of Data Analytics", <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2927540 > [accessed March 13, 2017]. See also Jonas Almeida, Ph.D. and others, "Big Data In Healthcare And Life Sciences Anonos Bigprivacy Technology Briefing", <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2941953> [accessed April 12, 2017].
.

obligations seriously would then make it possible to then assess the likelihood of the singling out, linkability and inference risks leading to re-identification and could make certain types of singling out, linking and inferring practices possible, as long as the purpose of the processing is not to re-identify data subjects and there is not a reasonable likelihood that the processing will lead to re-identification. It is true, nevertheless that the choice of confidentiality obligations coupled with weak sanitisation techniques can prove problematic if datasets are shared with multiple parties, even if each receiving party agrees to be bound by confidentiality obligations and adopt internal policies for this purpose. Obviously, access restrictions techniques and policies are a crucial means to make sure confidentiality obligations and policies are performed and/or implemented in practice.

Notably, while in the Breyer case of 2016 the CJEU interpreting the notion of "additional data which is necessary in order to identify the user of a website" considered the information held by the user's internet access provider, the CJEU recognised the importance of legal means in order to characterise personal data.[47] We suggest contractual obligations should be taken seriously into consideration in particular when they are backed up by technical measures such as measures to restrict access and dynamic measures to mitigate linkability.

[47] CJEU, C-582/14, Patrick Breyer v Bundesrepublik Deutschland, 19 October 2016, EU:C:2016:779. See in particular paragraph 39 where the CJEU, interpreting the DPD, states:

> Next, in order to determine whether, in the situation described in paragraph 37 of the present judgment, a dynamic IP address constitutes personal data within the meaning of Article 2(a) of Directive 96/45 in relation to an online media services provider, it must be ascertained whether such an IP address, registered by such a provider, may be treated as data relating to an 'identifiable natural person' where the additional data necessary in order to identify the user of a website that the services provider makes accessible to the public are held by that user's internet service provider.

4. Conclusion

The purpose of this paper was to test the possibility of interpreting the GDPR and Art. 29 WP's Opinion on Anonymisation Techniques together, assuming the concept of identifiability has two legs (identified and identifiable), the three risks of singling out, linkability and inference are relevant for determining whether an individual is identifiable and the concept of identifiability is used consistently across the GDPR. On the basis of an interdisciplinary methodology, this paper therefore builds a common terminology to describe different data states and derive the meaning of key concepts emerging from the GDPR: anonymised data, pseudonymised data and Art. 11 data. It then unfolds a risk-based approach, which is suggested to be compatible with the GDPR, by combining data sanitisation techniques and contextual controls in an attempt toeffectively balance data utility and data protection requirements. The proposed approach relies upon a granular analysis of re-identification risks expanding upon the threefold distinction suggested by Art. 29 WP in its Opinion on Anonymisation Techniques. It thus starts from the three common re-identification risks listed as relevant by Art. 29 WP, i.e. singling out, linkability and inference and further distinguishes between local, domain and global linkability to capture the key concepts of additional information and pseudonymisation introduced in the GDPR and comprehend the domain of Article 11 as well as the implications of Recital 26. Consequently, the paper aims to make it clear that even if a restrictive approach to re-identification is assumed, the GDPR makes the deployment of a risk-based approach possible: such an approach implies the combination of both contextual controls and sanitisation techniques and thereby the adoption of a relativist approach to data protection law. Among contextual controls, confidentiality obligations are crucial in order to reasonably mitigate re-identification risks.

Bibliography

Almeida, Jonas, Sean Clouston, Gary LaFever, Ted Myerson, and Sandeep Pulim, MD, "Big Data In Healthcare And Life Sciences Anonos Bigprivacy Technology Briefing", < https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2941953> [accessed April 12, 2017].

Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymisation Techniques, European Comm'n, Working Paper No. 216, 0829/14/EN (2014).

Article 29 Data Protection Working Party, Opinion 04/2007 on the concept of personal data, European Comm'n, Working Paper No. 136, 01248/07/EN (2007).

Byun, Ji-Won, Tiancheng Li, Elisa Bertino, Ninghui Li, and Yonglak Sohn, "Privacy-Preserving Incremental Data Dissemination", Journal Of Computer Security, 17/1 (2009): 43-68.

Dalenius, Tore, "Finding a needle in a haystack or identifying anonymous census records." *Journal of official statistics* 2, no. 3 (1986): 329.

Dwork, Cynthia, "Differential Privacy: A Survey Of Results", in *International Conference On Theory And Applications Of Models Of Computation* (Berlin Heidelberg, 2008): 1-19.

Hintze, Mike, "Viewing The GDPR Through A De-Identification Lens: A Tool For Clarification And Compliance", <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2909121> [accessed March 13, 2017].

Hintze, Mike, and Gary LaFever, "Meeting Upcoming GDPR Requirements While Maximizing The Full Value Of Data Analytics", <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2909121> [accessed March 13, 2017].

El Emam, Khaled, "Heuristics For De-Identifying Health Data." *IEEE Security & Privacy Magazine*, 6/4 (2008): 58-61.

El Emam, Khaled, Cecilia Álvarez, "A critical appraisal of the Article 29 Working Party Opinion 05/2014 on data anonymization techniques". *International Data Privacy Law* 2015, 5 (1): 73-87.

El Emam, Khaled, Eloise Gratton, Jules Polonetsky, Luk Arbuckle. "The Seven States of Data: When is Pseudonymous Data Not Personal Information?" <https://fpf.org/wp-content/uploads/2016/05/states-v19-1.pdf> [accessed March 13, 2017].

Information Commissioner's Office, *Anonymisation: Managing Data Protection Risk Code Of Practice*, 2012.

International Organization for Standardization, *ISO/TS 25237:2008 Health Informatics – Pseudonymization*, 2008 <https://www.iso.org/standard/42807.html> [accessed 13 March 2017].

Leibniz Institute for Educational Trajectories (LIfBi), Star Ng Cohort 6: Adults (SC6) SUF Version 7.0.0 Anonymiza On Procedures Tobias Koberg, 2009 <https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC6/7-0-0/SC6_7-0-0_Anonymization.pdf> [accessed 13 March 2017].

Machanavajjhala, Ashwin, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam, "L-Diversity", *ACM Transactions On Knowledge Discovery From Data*, 1/1 (2007).

Marianthi, Theoharidou, Spyros Kokolakis, Maria Karyda, and Evangelos Kiountouzis, "The Insider Threat To Information Systems And The Effectiveness Of ISO17799", *Computers & Security*, 24/6 (2005): 472-484.

Polonetsky, Jules, Omer Tene, and Kelsey Finch. "Shades of Gray: Seeing the Full Spectrum of Practical Data De-Identification." *56 , 3 Santa Clara Law Review* (2016): 593-629.

Ranjit Ganta, Srivatsava, Shiva Prasad Kasiviswanathan, and Adam Smith, "Composition Attacks And Auxiliary Information In Data Privacy", *Proceeding Of The 14Th ACM SIGKDD International Conference On Knowledge Discovery And Data Mining - KDD 08*, (2008).

Schwartz, Paul M., and Daniel J. Solove. "The PII problem: Privacy and a new concept of personally identifiable information." *New York University Law Review, Vol. 86* (2011): 1814.

Stalla-Bourdillon, Sophie, Alison Knight. "Anonymous data v. Personal data–A false debate: An EU perspective on anonymisation, pseudonymisation and personal data." *Wisconsin International Law Journal* (2017): 284-322.

Sweeney, Latanya, "K-Anonymity: A Model For Protecting Privacy", *International Journal Of Uncertainty, Fuzziness And Knowledge-Based Systems*, 10/05 (2002): 557-570.