ANONYMOUS DATA V. PERSONAL DATA—A FALSE DEBATE: AN EU PERSPECTIVE ON ANONYMIZATION, PSEUDONYMIZATION AND PERSONAL DATA

SOPHIE STALLA-BOURDILLON AND ALISON KNIGHT*

INTRODUCTION

This era of big data analytics promises many things. In particular, it offers opportunities to extract hidden value from unstructured raw datasets through novel reuse. The reuse of personal

* Sophie is Associate Professor in Information Technology Law at the University of Southampton (UK) and Director of the Institute for Law and the Web. She is blogmaster at https://peepbeep.wordpress.com/.

Alison Knight is a Research Fellow in Law, and member of the Web Science Institute, at the University of Southampton. She has previously worked for US and UK law firms in Brussels and London, as well as the UK Government Legal Service.

Electronic copy available at: https://ssrn.com/abstract=2927945

data is, however, a key concern for data protection law as it involves processing for purposes beyond those that justified its original collection, at odds with the principle of purpose limitation.

The issue becomes one of balancing the private interests of individuals and realizing the promise of big data. One way to resolve this issue is to transform the personal data that will be shared for further processing, such as data mining, into ―anonymous information‖ to use an EU legal term. ―Anonymous information‖ is outside the scope of data protection laws in the EU,[1] and is also carved out from privacy laws in many other jurisdictions worldwide.

The foregoing solution works well in theory, but only as long as the output potential from the data still retains utility, which is not necessarily the case in practice. This is because the value or knowledge that can be gained from analyzing datasets (particularly using automatic algorithmic software) is maximized by virtue of finding patterns, basically linking relationships between data points. Anonymization, by contrast, aims to delink such data point relationships where they relate to informational knowledge that can be gleaned in respect of specific persons and their identities. This leaves those in charge of processing the data with a problem: how can they ensure that anonymization is conducted effectively on the data on their possession, while retaining that data‘s utility for potential future disclosure to, and further processing by, a third party?

Despite a broad consensus around the need for effective anonymization techniques, the debate as to when data can be said to be legally anonymized to satisfy EU data protection laws is a long-standing one. Part of the complexity in reaching consensus on this issue derives from confusion around terminology, in particular the meaning of the concept of anonymization in this context, and how strictly delineated that concept should be. This can be explained, in turn, by a lack of consensus

[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]; 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].

Electronic copy available at: https://ssrn.com/abstract=2927945

on the doctrinal theory that should underpin its traditional conceptualization as a privacy-protecting mechanism.

For example, the texts of both the existing EU Data Protection Directive[2] (DPD) and the new EU General Data Protection Regulation[3] (GDPR) are ambiguous. Although both legal instruments seem to adopt a restrictive definition of ―anonymous information,‖ to mean anonymous in such a way that the data subject is no longer identifiable, they also seem to support a risk-based approach that limits the identifiable concept by a reasonableness standard, i.e. to the extent that only all the means likely reasonably to be used to identify someone are taken into account.[4]

In fact, the GDPR seems even more restrictive than the DPD, because it introduces a new category of data (i.e. data that has undergone pseudonymization)[5]—which is distinguished from the category of ―anonymous information,‖ or data that has undergone anonymization to use better terminology.[6] The concept of ―pseudonymisation‖ is defined under the GDPR to mean:

> [T]he 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.[7]

Personal data that has undergone pseudonymization is explicitly defined to remain personal data under EU data protection laws, as it

[2] DPD, *supra* note 1.

[3] The GDPR was agreed by the European Commission, European Parliament, and the Council of the EU in December 2015 to replace the DPD. In April 2016, the European Parliament formally approved the final text version of the GDPR for translation into the EU‘s official languages. It came into force on 24 May 2016 and takes effect on 25 May 2018. Thus, organizations have two years within which to ensure that they comply with the GDPR in anticipation of that date. GDPR, *supra* note 1.

[4] DPD, *supra* note 1, at Recital 26; GDPR, *supra* note 1, at Recital 26.

[5] GDPR, *supra* note 1, at Recital 26 and Article 4(5).

[6] We will, however, use the term ―anonymized data‖ in this article as it is shorter and makes the reading easier. We are nonetheless of the view that the expression ―data that has undergone anonymization‖ better captures the idea of data characteristics as fluid concepts which, as a matter of fact, can only be understood in the context of appreciating ongoing processes related to the data environment, and which does not ‗simply‘ focus upon data as having static and immovable qualities as we will explain below. A similar choice has been made by others. *See, e.g.,* MARK ELLIOT ET AL., THE ANONYMISATION DECISION-MAKING FRAMEWORK 1 (Ukan Publ‘ns, 2016 ed.).

[7] GDPR, *supra* note 1, at Article 4(5).

―should be considered to be information on an identifiable natural person.‖[8]

This paper suggests that although the concept of anonymization is crucial to demarcate the scope of data protection laws at least from a descriptive standpoint, recent attempts to clarify the terms of the dichotomy between ―anonymous information‖ and personal data (in particular, by data protection regulators in the EU) have partly failed. Although this failure could be attributed to the very use of a terminology that creates the illusion of a definitive and permanent contour that clearly delineates the scope of data protection laws, the reasons for such a failure are slightly more complex. Essentially, this failure can be explained by the implicit adoption of a static approach, which tends to assume that once the data is anonymized, not only can the initial data controller forget about it, but also the recipients of the dataset are free from any obligation or duty because the transformed dataset always lies outside the scope of data protection laws. By contrast, the state of anonymized data has to be comprehended in context, which includes an assessment of the data, the infrastructure, and the agents.[9] Moreover, it is very important to comprehend the state of anonymized data dynamically. This dynamic state is epitomized by the fact that anonymized data can become personal data again, depending upon the purpose of the further processing and future data linkages.

The implications of this dynamic approach are, in particular, that recipients of anonymized data, although they are not data controllers

[8] To note, while the final GDPR text does not make pseudonymous data (so defined) a special category of personal data - in the sense that it does not seem to benefit from a light-touch data protection regime in being exempted from certain data protection rules, doubts still persist over the exact implications of its status. For example, see the formulation of Recital 29 of the GDPR, which states:

> ‒[i]n order to create incentives for applying pseudonymisation when processing personal data, measures of pseudonymisation whilst allowing general analysis should be possible within the same controller when the controller has taken technical and organizational measures necessary to ensure, for the respective processing, that the provisions of this Regulation are implemented, and ensuring that additional information for attributing the personal data to a specific data subject is kept separately.‖ *Id.* at Recital 29.

Besides Article 6(4) of the GDPR provides that in order to ascertain whether further processing is compatible with the purpose of the initial processing, considerations relating to the existence of appropriate safeguards such as encryption or ‒pseudonymisation‖ should be taken into account. *Id.* at Article 6(4).

[9] ELLIOT ET AL, *supra* note 6, at 2 (‒The framework is underpinned by a relatively new way of thinking about the re-identification problem which posits that you must look at both the data and the data environment to ascertain realistic measures of risk.‖).

when they receive the dataset, have to behave responsibly and comply with any licensing obligations imposed by the original data controllers of the raw personal data. Specifically, the former must abide by any licensing limitations upon the purpose and the means of the processing of the data in its disclosed post-anonymization process form to remain outside the scope of data protection laws. At the same time, the characterization of anonymized data should also be dependent upon an ongoing monitoring on the part of the initial data controller of the data environment of the dataset that has undergone anonymization.

The paper starts by examining the recent approaches to anonymization, highlighting in particular the shortcomings of the legal and technical approaches to this issue adopted at the EU level, which are based implicitly on a static approach, assuming that once the anonymized dataset is released the recipient has complete freedom of use over its subsequent processing. The paper then unfolds the main components of a dynamic approach, and explains why an approach to anonymization of this type is both more appropriate and compatible with the current and soon-to-be-applied EU legal framework under the GDPR. Ultimately, the paper makes the point that the opposition between so-called ―anonymous information‖ and personal data in a legal sense is less radical than usually described.

I. THE SHORTCOMINGS OF RECENT LEGAL AND TECHNICAL APPROACHES TO THE CONCEPT OF ANONYMIZATION

While the DPD was adopted relatively early in 1995, somewhat surprisingly, prior to 2014, there was no comprehensive guidance interpreting and ‗unpacking' the DPD's provisions on anonymization at the EU level. That changed with the release of an Opinion by the Article 29 Data Protection Working Party (―Art. 29 WP‖) on ―Anonymisation Techniques‖[10] (―Anonymization Opinion‖). The opinion covers a range of legal and technical issues surrounding data anonymization. It is not a step-by-step manual on how to go about anonymization. Rather, it sets out a good practice framework to enable data controllers to make better decisions about carrying out anonymization.

[10] *See* 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*].

Art. 29 WP‘s Anonymization Opinion was released two years after the release of the Code of Practice on ―Anonymisation: Managing Data Protection Risks‖ (―the Code‖) by the UK Information Commissioner‘s Office (ICO), the UK data protection agency. Despite the fact that Art. 29 WP comprises representatives of the data protection supervisory authorities designated by each EU Member State, including the UK, its opinions are not always consensual. This is the case of the Anonymization Opinion, which departs from the ICO‘s Code on a significant point, as will be explained.

Thus, it is useful to discuss the different sources of law, be it hard or soft, in the order they were adopted. The analysis will start with a brief description of the DPD, and follow with consideration of the approach adopted by the ICO in the Code in 2012, thereafter to continue with an explanation of the position adopted by Art. 29 WP in 2014. The analysis will then examine whether the final text of the GDPR made public in April 2016 and published in the EU Official Journal on 4 May 2016 implicitly endorses the approach taken in the DPD to anonymization and the interpretations thereto by the data protection authorities, or adopts a new approach. Finally, we will quickly mention the data sharing initiative presented by the UK Cabinet Office in early 2016, as well as the Big Data report of the EU Agency for Network and Information Security (ENISA), released in December 2015, to further stress the importance of accurate and workable definitions in this field.

A. THE DPD

For background, under the DPD, a data controller is required to justify the processing of personal data before it will be considered lawful under EU data protection laws. Article 7 sets out a number of legitimizing grounds for processing with which data controllers must comply.[11] The most popular justification is to obtain the unambiguous

[11] Member States shall provide that personal data may be processed only if:

> (a) the data subject has unambiguously given his consent; or
>
> (b) processing is necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract; or
>
> (c) processing is necessary for compliance with a legal obligation to which the controller is subject; or
>
> (d) processing is necessary in order to protect the vital interests of the data subject; or

consent of the data subject to the processing under Article 7(a); the other conditions can be found in Article 7(b)-(f). Article 6[12] of the DPD, furthermore, sets out a number of principles with which data controllers must comply when processing personal data. In particular, the principles oblige the data controller to process the data fairly and lawfully in pursuance of Article 6(1)(a), and to collect data only for specified, explicit and legitimate purposes, and not to further process it in any manner incompatible with those purposes, also known as the ―purpose limitation‖ principle as per Article 6(1)(b). If there is incompatibility with the initial processing purpose, a new legitimizing ground for the further processing must generally be found, although it will not save the further processing in all cases.[13]

(e) processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller or in a third party to whom the data are disclosed; or

(f) processing is necessary for the purposes of the legitimate interests pursued by the controller or by the third party or parties to whom the data are disclosed, except where such interests are overridden by the interests for fundamental rights and freedoms of the data subject which require protection under Article 1 (1). DPD, supra note 1, at Article 7.

[12] 1. Member States shall provide that personal data must be:

(a) processed fairly and lawfully;

(b) collected for specified, explicit and legitimate purposes and not further processed in a way incompatible with those purposes. Further processing of data for historical, statistical or scientific purposes shall not be considered as incompatible provided that Member States provide appropriate safeguards;

(c) adequate, relevant and not excessive in relation to the purposes for which they are collected and/or further processed;

(d) accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that data which are inaccurate or incomplete, having regard to the purposes for which they were collected or for which they are further processed, are erased or rectified;

(e) kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the data were collected or for which they are further processed. Member States shall lay down appropriate safeguards for personal data stored for longer periods for historical, statistical or scientific use.

2. It shall be for the controller to ensure that paragraph 1 is complied with.

DPD, *supra* note 1, at Article 6.

[13] *See* GDPR, *supra* note 1, at Recital 40 (confirms position); EUROPEAN UNION AGENCY FOR FUNDAMENTAL RIGHTS & COUNCIL OF EUR., HANDBOOK ON EUROPEAN DATA PROTECTION LAW 70 (2014). With that said, there does not seem to be consensus on this ground as Art. 29 WP appears to be of the opinion that even if the purpose of the further processing is compatible with the purpose of the initial processing a new legal basis is required. *See* Article 29 Data Protection Working Party, *Opinion 03/2013 on Purpose Limitation* (European Comm'n, Working Paper No. 203, 00569/13/EN, 2013), at 27.

Personal data is defined by the DPD as reading any information relating to an identified or identifiable natural person (―data subject‖).[14] Article 2(a) provides:

> ‗personal data‘ shall mean any information relating to an identified or identifiable natural person (‗data subject‘); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity.[15]

In the preamble to the DPD, Recital 26 also reads as follows:

> Whereas the principles of protection must apply to any information concerning an identified or identifiable person; whereas, to determine whether a person is identifiable, account should be taken of all the means likely reasonably to be used either by the controller or by any other person to identify the said person; whereas the principles of protection shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable; whereas codes of conduct within the meaning of Article 27 may be a useful instrument for providing guidance as to the ways in which data may be rendered anonymous and retained in a form in which identification of the data subject is no longer possible.[16]

In other words, although Article 2(a) suggests a very wide scope to the legal definition of personal data, the non-binding, but highly persuasive interpretation of Article 2(a) in Recital 26 of the DPD, appears to limit this definition using a ―means likely reasonably‖ standard.[17] Going further, the DPD in Recital 26 appears to adopt a risk-based approach to the definition of personal data, and, thereby, to the legal effects of anonymization processes. While the ―data [is] rendered anonymous‖ if and only if the ―data subject is no longer identifiable,‖ the reversibility of the de-identification process should not mean that the data can never fall outside the scope of data protection law.[18] To determine whether the data is (legally) rendered anonymized, it is enough to assess (and to some extent anticipate) ―the means likely reasonably to be used‖ by the data controller and third parties by which they could re-identify the data subject.[19]

[14] DPD, *supra* note 1, at Article 2(a).
[15] *Id.*
[16] DPD, supra note 1, at Recital 26.
[17] DPD, supra note 1, at Recital 26.
[18] DPD, supra note 1, at Recital 26.
[19] DPD, supra note 1, at Recital 26.

B. THE UK ICO‗S CODE OF PRACTICE ON ANONYMIZATION: MANAGING DATA PROTECTION RISK

The Code is aimed at data controllers in public, private, and third-sector organizations, who use anonymization techniques. Although the Code does not have the force of law, the UK Information Commissioner has stated that he will consider the Code when investigating anonymization-related issues.[20]

Like the Art. 29 WP‗s Anonymization Opinion to be discussed in the next sub-section, the Code provides practical advice on the methods for anonymizing data and the associated risks of publishing such data. In particular, it includes guidance on assessing the risk of re-identification—for example through data linkage, i.e. matching data points through unique patterns.

The Code stresses that organizations must evaluate whether an individual can be identified from an anonymized dataset, either by itself, or through data linkage in combination with other information that might be available.[21] In this respect, the Code stresses the need to evaluate the likely availability of such ―other information‖[22] to a third party (by which they might be able to re-identify an individual).[23] In this sense, the ICO does not adopt an identifiability assessment approach that is only concerned about the viewpoint of the data controller, and the means likely reasonably to be used by the data controller in achieving identification of an individual from the data. Notably, and this is really important, this is true even if the definition of personal data found in section 1(1) of the UK Data Protection Act (1998) (DPA) appears more restrictive than the definition contained in the DPD. Section 1(1) states that ―personal data‖ means ―data which relate to a living individual who can be identified – (a) from those data, or, (b) from those data and other

[20] INFO. COMM‗R‗S OFFICE, ANONYMISATION: MANAGING DATA PROTECTION RISK CODE OF PRACTICE 7 (2012), https://ico.org.uk/media/1061/anonymisation-code.pdf [hereinafter CODE OF PRACTICE]; *See also* INFO. COMM‗R‗S OFFICE, CODE OF PRACTICE FOR THE SHARING OF PERSONAL INFORMATION (2011), https://ico.org.uk/media/1068/data_sharing_code_of_practice.pdf. The latter Code sets out a model of good practice for public, private and third-sector organizations, and covers systematic, routine data-sharing where the same data sets are shared between the same organizations for an established purpose, as well as one-off instances where a decision is made to release data to a third party. In particular, it includes public and private sector case studies to explain practically how the UK Data Protection Act (1998) applies to data-sharing arrangements.

[21] CODE OF PRACTICE, *supra* note 20, at 17–18.

[22] CODE OF PRACTICE, *supra* note 20, at 18.

[23] CODE OF PRACTICE, *supra* note 20, at 17-18.

information which is in the possession of, or is likely to come into the possession of, the data controller . . .‖[24] Thus, in describing ways of assessing re-identification risk, the Code emphasizes that the identification capabilities of third parties—in associating the data with ―other information‖ that might allow them to re-identify a particular individual post-anonymization—must be considered in tandem with the identification capabilities of the data controller.[25] Notwithstanding, the ICO admits that this task is ―extremely problematic‖ as ―more potentially ‗match-able‘ information becomes publicly available.‖[26] Given the principle of supremacy of EU law and the duty of consistent interpretation developed by the Court of Justice of the European Union (CJEU),[27] such a liberal interpretation aligns itself with the approach taken under the DPD (Recital 26) and is clearly sensible.

The Code recommends adopting a ―motivated intruder‖ test as an aid to assessment.[28] This term refers to a hypothetical entity ―who starts without any prior knowledge but who wishes to identify the individual from whose personal data the anonymised data has been

[24] To note, the importance to be placed upon the data controller‘s identifiability perspective was recently considered by the English Court of Appeals decision in Google Inc. v. Judith Vidal-Hall and others, in which the Court considered Recital 26 of the DPD. Google Inc. v. Vidal-Hall [2015] EWCA (Civ) 311, at ¶ 105 (U.K.). The Court stated that it was arguable that the test to be found in Recital 26 (including both the data controller and third parties) was to prevail over Article 2(a) of the DPD which does not expressly refer to third parties to define personal data and identifiable data subjects. *Id.* at ¶ 125. The Court was asked in preliminary proceedings to consider claims against Google for installing cookies on Apple‘s ‗Safari‘ internet browser without specific users‘ consent, specifically to enable advertising to be targeted at users. Google first argued that as it was not ―reasonably likely‖ that it would aggregate two sets of data in its possession to identify such users from browser-generated information (BGI) it had collected about them, so such BGI could not be deemed personal data under the DPA. Then, Google said that, even if the Safari browser users were identifiable to others, it did not matter, as the data controller‘s perspective is the only one that matters under the DPA. In other words, Google argued that the potential identification of particular individuals to notional third parties (from targeted advertising inherently revealing the BGI to those with sight of the users‘ device screens) was an impermissible route to identification under UK law. The Court rebutted these arguments insofar as it agreed that their counter-arguments should be tested at full trial. *Id.*

[25] Accordingly, we submit that hints by the European Commission that the UK failed in the DPA to implement the DPD adequately in this respect could now be appeased. *See, e.g.*, DOUWE KORFF, EC STUDY ON IMPLEMENTATION OF DATA PROTECTION DIRECTIVE 95/46/EC: REPORT ON THE FINDINGS OF THE STUDY 13 (Human Rights Ctr. 2002), http://194.242.234.211/documents/10160/10704/Stato+di+attuazione+della+Direttiva+95-46- CE.

[26] CODE OF PRACTICE, *supra* note 20, at 18.

[27] Case C-106/89, Marleasing SA v. La Comercial Internacionale de Alimentacion SA, 13 Novembre 1990, ECLI:EU:C:1990:395 (EU).

[28] CODE OF PRACTICE, *supra* note 20, at 22.

derived.‖[29] It is proposed as a tool for assessing re-identification risk. The ICO extends its analysis by encouraging consideration of whether such a hypothetical attacker would be able to identify an individual, assuming they have certain motivation, means, and skills attributed to them. These include access to resources (such as libraries and the Internet), in addition to the ability to employ investigative techniques, such as ―making enquiries of people who may have additional knowledge of the identity of the data subject or advertising for anyone with information to come forward.‖[30] In other words, the _motivated intruder' test sets the bar higher than that of a relatively inexpert member of the public attempting data re-identification. Despite being attributed with reasonable competence, however, the attacker is not presumed to have any ―specialist knowledge such as computer hacking skills, or to have access to specialist equipment or to resort to criminality such as burglary, to gain access to data that is kept securely.‖[31]

To this end, the Code suggests through its analysis that if an organization takes reasonable security and disclosure limitation steps in respect to data that has been subject to anonymization techniques, the subsequent processing of that data should not necessarily be caught by the DPA.[32] (Ultimately, whether it is caught will depend on an assessment of the means likely reasonably standard as applied to the circumstances at issue). The Code further highlights the importance of secondary factors in informing an organization's approach to data disclosure and security, such as the type of data at issue and who it is about, both factors of which could impact upon the implications for an individual in case they are re-identified. The Code states:

> Clearly, some sorts of data will be more attractive to a _motivated intruder' than others. Obvious sources of attraction to an intruder might include: finding out personal data about someone else, for nefarious personal reasons or financial gain; the possibility of causing mischief by embarrassing others; revealing newsworthy information about public figures; political or activistic purposes, eg as part of a campaign against a particular organisation or person; or curiosity, eg a local person's desire to find out who has been involved in an incident shown on a crime map.[33]

[29] CODE OF PRACTICE, *supra* note 20, at 22.

[30] CODE OF PRACTICE, *supra* note 20, at 22-23.

[31] CODE OF PRACTICE, *supra* note 20, at 23.

[32] CODE OF PRACTICE, *supra* note 20, at 13.

[33] CODE OF PRACTICE, *supra* note 20, at 23.

The ICO also takes the opportunity, in the Code, to distinguish anonymization—that is, data where no information relating to or identifying any individual is shown—as a process dependent on data aggregation from a process that depends upon removing certain individual identifiers from person-specific data, which leaves individual-level data in the dataset. The latter, it says, carries higher risks, although not insurmountable ones to effective anonymization. The ICO states in the Code:

> [E]ven though pseudonymised data does not identify an individual, in the hands of those who do not have access to the ‗key', the possibility of linking several anonymised datasets to the same individual can be a precursor to identification. This does not mean though, that effective anonymisation through pseudonymisation becomes impossible.[34]

Notably, the ICO defines pseudonymization as ―[t]he process of distinguishing individuals in a dataset by using a unique identifier which does not reveal their ‗real world' identity.‖[35] Given the approximation of the definition of pseudonymization, as using a unique identifier does not necessarily mean that real world identities are not revealed, it is not entirely clear reading the Code which additional steps would be required to pass from pseudonymized data to anonymized data.

With that said, the foregoing does suggest that the ICO interprets anonymization as a process that can be good enough depending on the circumstantial facts (rather than as an absolute state), even if the ICO does not fully explain how the anonymization through pseudonymization is possible. In other words, the foregoing suggests that the ICO adopts a risk-based approach to anonymized data.

C. ART. 29 WP'S OPINION ON ANONYMIZATION TECHNIQUES

Like the ICO before it, Art. 29 WP combines both a legal and a technical approach to anonymization. It is technical in the sense that it explores the state of the art in the field focusing on current technologies used to pursue anonymization objectives. For example, the WP discusses randomization and generalization, noise addition, permutation, differential privacy, aggregation, k-anonymity, l-diversity, and t-

[34] CODE OF PRACTICE, *supra* note 20, at 21.

[35] CODE OF PRACTICE, *supra* note 20, at 29, 49.

closeness. Strengths and weaknesses of techniques are highlighted along with common mistakes and failures.

The Anonymization Opinion also makes recommendations on using the techniques in the light of the risk of re-identification of individuals, which should assist data controllers with designing an anonymization process. Furthermore, it clarifies that pseudonymization, which Art. 29 WP describes as a process by which one attribute—typically a unique one—in a record is replaced for another, is not a method of anonymization, but merely a useful security measure.[36] In this sense, the definition of pseudonymization put forward by Art. 29 WP in its opinion is more helpful than the ICO‘s definition.

To understand why, it is useful to consider what the Anonymization Opinion says about the nature of re-identification risk and its recommendations on using such methods in the light of different categories of the risk of identification of individuals. In particular, the Opinion treats the robustness of anonymization techniques against re-identification, ―performed by the most likely and reasonable means the data controller and any third party may employ,‖ based upon three different types of risk:

1. ***Is it still possible to single out an individual?*** (― . . . which corresponds to the possibility to isolate some or all records which identify an individual in the dataset‖);[37] or,

2. ***Is it still possible to link records relating to an individual?*** (― . . . 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)‖);[38] or,

3. ***Can information be inferred concerning an individual?*** (― . . . which is the possibility to deduce, with significant probability, the value of an attribute from the values of a set of other attributes‖).[39]

Therefore, for Art. 29 WP, pseudonymized data remains personal data, the processing of which remains subject to data protection law, because at least one of these risk categories cannot be excluded. In other words,

[36] *Opinion on Anonymisation Techniques*, *supra* note 10, at 20.

[37] *Opinion on Anonymisation Techniques*, *supra* note 10, at 11.

[38] *Opinion on Anonymisation Techniques*, *supra* note 10, at 11.

[39] *Opinion on Anonymisation Techniques*, *supra* note 10, at 12. To this extent, it would appear that Art. 29 WP is setting the regulatory bar for achieving anonymization in law higher than that of the ICO.

while pseudonymization reduces the linkability of a dataset with the original identity of a data subject, meaning the processing of pseudonymized data involves fewer risks for the individual, it does not necessarily reduce it significantly.

In the Anonmyisation Opinion, Art. 29 WP does include statements that suggest that it is sympathetic to a risk-based approach. For example, Art.29 WP warns of the difficulty of creating ―a truly anonymous dataset‖ where much underlying information is retained, as combining an anonymized dataset with another dataset may lead to identification.[40] Like the Code before it, strengths and weaknesses of anonymization techniques are highlighted, along with recommendations for improvement in the light of re-identification risk. The Opinion also encourages consideration of an aid to assessing re-identification risk by postulating a possible attacker trying to de-identify data for their own purposes.[41] In this vein, Art. 29 WP emphasizes the need to take into account contextual elements, such as the nature of the data and any informational disclosure control mechanisms in place to restrict access to the data.[42] The Anonymization Opinion also advises that anonymization should be planned on a case-by-case basis, possibly using a variety of techniques and factoring in the Opinion‘s recommendations.[43] Thus, data controllers are advised not to treat anonymization as a one-off exercise. Rather, regular risk assessment should continue in the light of the residual risk of identification.[44]

Art. 29 WP‘s position remains problematic, however, because of statements found elsewhere in the Anonmyisation Opinion. Tellingly, the Anonymisation Opinion states that each of the techniques tested ―fails to meet with certainty the criteria of effective anonymisation.‖[45] Besides, it affirms that it interprets ―anonymised data‖ to mean ―anonymous data that previously referred to an identifiable person, but where that identification is no longer possible.‖[46] While presenting the technical issues and risks inherent to anonymization, therefore, this statement

[40] *Id.* at 3.

[41] *Opinion on Anonymisation Techniques*, *supra* note 10, at 11–12.

[42] *Opinion on Anonymisation Techniques*, *supra* note 10, at 25.

[43] *Opinion on Anonymisation Techniques*, *supra* note 10, at 23-25.

[44] *Opinion on Anonymisation Techniques*, *supra* note 10, at 4.

[45] *Opinion on Anonymisation Techniques*, *supra* note 10, at 23.

[46] *Opinion on Anonymisation Techniques*, *supra* note 10, at 6, 8. The Anonymization Opinion also alludes to ―irreversibility of the alteration undergone by personal data to enable direct or indirect identification‖ as key to definitions of anonymization in international standards (at 6).

suggests that Art. 29 WP's understanding of acceptable re-identification risk requires near-zero probability, an idealistic and impractical standard that cannot be guaranteed in a big data era.

One even finds the adjective ―irreversible‖ to describe the anonymization process a few paragraphs earlier:

> More precisely, the data must be processed in such a way that it can no longer be used to identify a natural person by using ―all the means likely reasonably to be used‖ by either the controller or a third party. An important factor is that the processing must be irreversible.[47]

With this said, the truly problematic part of the Anonmyisation Opinion is contained in this sentence:[48]

> [I]t is critical to understand that when a data controller does not delete the original (identifiable) data at event-level, and the data controller hands over part of this dataset (for example after removal or masking of identifiable data), the resulting dataset is still personal data.[49]

Art. 29 WP is thus stating that transformed data, or data that has passed through an anonymization process, can never amount to ―data rendered anonymised‖ within the meaning of EU data protection law so long as the initial raw dataset comprising information about identified or identifiable data subjects has not been destroyed by the data controller.[50]

By affirming such statements, and despite what is says elsewhere in the same Opinion, Art. 29 WP appears to reject the very consequences of a risk-based approach. This is because, if it is possible to isolate the raw datasets from the transformed datasets and put in place security measures, including technical and organizational measures, as well as legal obligations (essentially contractual obligations) so that the subsequent recipient of the transformed dataset will never have access to the raw dataset, the transformed dataset should be deemed as comprising data rendered anonymous at the very least in the hands of the subsequent recipient of the dataset. This is all the more warranted, as we will explain below, because where the transformed dataset is still considered to be

[47] *Id.* at 5.

[48] Francis Aldhouse, Former Deputy Info. Comm'r, Keynote Address at the University of Southampton Seminar on Defining and Regulating Anonymisation (Mar. 9, 2016), www.southampton.ac.uk/ilaws/news/events/2016/03/09-defining-and-regulating-anonymisation.page.

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

[50] CODE OF PRACTICE, *supra* note 20, at 13. *See also* Common Servs. Agency v. Scottish Info. Comm'r [2008] UKHL 47, at ¶¶ 27, 92 (appeal taken from Scot.).

personal data in the hands of the subsequent recipient of the dataset, complying with the entire gamut of the data protection obligations is likely to have a chilling effect on data sharing given the complexity of distilling the new compliance obligations that will be required to be followed by organizations under the GDPR. This could have, in particular, a significant deterrent effect on the carrying out of beneficial longitudinal research studies that rely upon the reuse of personal data once it has undergone anonymization methods such as in the health or education sectors.

D. THE GDPR

The GDPR defines personal data in Article 4(1) as:

> any information relating to an identified or identifiable natural person (‗data subject‘); an identifiable 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.

This is essentially the same as the DPD definition, save additional list of criteria by which personal data can be identified. As a result, the legal trend in the EU is to not reduce the scope of the category of personal data.

The expansive scope of personal data under EU law is confirmed by other changes, and in particular, the introduction of the new concept of identification as encompassing ―singling out‖[51] in Recital 26 of the GDPR. The language adopted in this Recital 26, however, also makes it clear that under the new regime, data protection principles will continue not to apply to anonymized data. To use the exact terms of the GDPR (our emphasis):

> To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing

[51] GDPR, *supra* note 1, at Recital 26.

> and technological developments. The principles of data protection should therefore not apply to anonymous information, namely information which 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. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes.[52]

It thus appears that the GDPR still adopts, at least in its recital, a risk-based approach to anonymization, relying upon the test of the ―means reasonably likely to be used‖ by the data controller and third parties, which as aforementioned comes from Recital 26 of the DPD.[53]

This being said, the GDPR goes beyond the DPD by introducing a new definition: that of ―pseudonymisation‖. As mentioned above, Article 4(5) of the GDPR reads as follow:

> ‗[P]seudonymisation' 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.[54]

Very importantly, this definition is both narrow and very broad. It is narrow in the sense that it excludes processes that cannot ensure that the personal data is not attributed to an identifiable natural person and this should be welcome. As a result, it is unlikely that data processed for online targeted advertising purposes will, at least legally, continue to be labelled as non-personal data that falls outside the scope of EU data protection laws or even as pseudonymized data. This is because, given the richness of the information collected in this context, and above all given the purpose of the processing, which is to evaluate the behavior of internet users, the data collected can still be deemed attributed to an identifiable, if not identified, natural person.

More problematically, however, the definition of ―pseudonymisation‖ is also very broad in the sense that, because there is no reference to data linkability as the inherent problem that belies concern that individuals may yet be singled out from data that has

[52] GDPR, *supra* note 1, at Recital 26.

[53] DPD, *supra* note 1, at Recital 26. Albeit to note that recitals of EU directives are not deemed binding law, but highly persuasive interpretations of provisions in the main body of a directive to follow.

[54] GDPR, *supra* note 1, at Article 4(5).

undergone ―pseudonymisation,‖ it could include data that has undergone aggregation practices to remove individual-level elements within it. This, we suggest, should not be welcomed.

To understand why the GDPR may be deemed to adopt such a broad definition of ―pseudonymisation,‖ we revert to the second sentence in Recital 26 of the GDPR: ―Personal data which has undergone pseudonymisation, which could be attributed to a natural person by the use of additional information, should be considered as information on an identifiable natural person.‖ One way to make sense of this sentence would be to say that, as long as the raw dataset has not been destroyed, a transformed dataset must only be considered pseudonymized and remain subject to EU data protection laws. The GDPR would thus be endorsing Art. 29 WP‘s approach to anonymization, which as explained above, is not fully consistent with a risk-based approach to anonymization.

A more nuanced interpretation of this sentence, building on the approach adopted by the ICO in the Code, by contrast, is that if anonymization through pseudonymization seems to fall short legally, there still remains a route to effective anonymization through aggregation. This interpretation makes better legal sense as the removal of individual-level elements within a shared dataset truncates in principle the possibility of any harm befalling to individuals through the linking of individualized data records from which they could be singling out. We suggest that this is the most sensible path despite the lack of reference to linkability in the definition to be found in Article 4 of the GDPR. As aforementioned, the lack of reference to linkability is problematic since as the definition of pseudonymization refers to both identified and identifiable data subjects the risk remains that data will be considered pseudonymized as long as the raw dataset has not been destroyed, even if the route of anonymization through aggregation has been chosen.

Pushing the reasoning further, there is another reason why the definition of pseudonymization to be found in the GDPR is problematic. Comparing Articles 4 and 11 of the GDPR,[55] it appears that de-identified

[55] Processing which does not require identification.

> 1. If the purposes for which a controller processes personal data do not or do no longer require the identification of a data subject by the controller, the controller shall not be obliged to maintain, acquire or process additional information in order to identify the data subject for the sole purpose of complying with this Regulation.
>
> 2. Where, in cases referred to in paragraph 1 of this Article, the controller is able to demonstrate that it is not in a position to identify the data subject, the controller shall

data in the sense of Article 11 (i.e. data for which the data controller —is not in a position to identify the data subject‖[56]) expressly benefits from a light-touch regime in the sense that, as a matter of principle, Articles 15 to 20 (i.e. data subject rights to access, rectification, erasure, restrictions, notification and data portability) do not apply, whereas nothing is expressly said about pseudonymized data. Notably, to determine whether data has undergone pseudonymisation the means reasonably likely to be used by third parties should be taken into account on top of the means reasonably likely to be used by the data controller since individuals shall not be identifiable. It seems nevertheless crucial not to always stop at the characterization of de-identified data within the meaning of Article 11, as pseudonymization is conceived as a means to effectuate the principle of data protection by design under Article 25 and the violation of such a principle could have severe consequences as per Article 83.

E. THE UK CABINET OFFICE CONSULTATION ON DATA SHARING

As mentioned above, a consultation was launched in February 2016[57] by the UK Cabinet Office specifically focused on enabling more data sharing between public sector organizations.[58] In seeking feedback on how the UK Government can use data to improve public services for citizens and to improve decision-making, it hoped to —maximise opportunities for effective data sharing and build on areas of good practice.‖[59] The consultation was also designed to assess how data is accessed and used by these organizations with an eye to improvement.

As mentioned, a key area on which the Government sought consultation views is in relation to what it terms —de-identified data‖ and official statistics to be shared with, and used by, researchers in order to

inform the data subject accordingly, if possible. In such cases, Articles 15 to 20 shall not apply except where the data subject, for the purpose of exercising his or her rights under those articles, provides additional information enabling his or her identification. GDPR, *supra* note 1, at Article 11.

[56] GDPR, *supra* note 1, at Article 11(2).

[57] U.K. Cabinet Office, *Better use of data in government: consultation outcome* (July 6, 2016), https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/535063/better_use_of_data_in_government_response_final.pdf.

[58] *Id.* at 29.

[59] Press Release, Cabinet Office and The Rt. Hon. Matt Hancock MP, Launch of new data sharing consultation (Feb. 29, 2016), https://www.gov.uk/government/news/launch-of-new-data-sharing-consultation.

carry out research for public benefit, e.g. in order to improve the quality of migration and population statistics.

For the purpose of this Consultation Paper, the key term of ―de-identified data‖ was defined as:

> [D]ata that does not directly identify a living individual, and so does not amount to ‗personal data' under the first limb of ‗personal data' under the Data Protection Act. This data could nonetheless potentially amount to personal data under the second limb of the definition if the individual to which it relates could be identified from the combination of that data with other data held or likely to be held by the data controller.[60]

This definition was juxtaposed on the same page of the Consultation Paper with a definition of ‒anonymisation‖ as ‒a process of rendering data into a form that does not identify individuals and where identification is not likely to take place.‖[61]

Reading these definitions in the light of the DPA, but also, and more importantly in the light of the GDPR, it would seem that de-identified data could still be personal data. Moreover, de-identified data is not necessarily equivalent to the category of data that has undergone a process of pseudonymization within the meaning of the GDPR because pseudonymization within the meaning of the GDPR shall ensure that the data cannot be attributed to an identifiable data subject.[62] This being said, the Cabinet Office is nonetheless of the view that de-identification is a satisfactory security measure given the purpose of the further processing of the data to support accredited researchers to access and link data in secure facilities to carry out research for public benefit.[63]

Therefore, this further processing of de-identified—yet potentially still legally personal—data is arguably considered by the Cabinet Office to be compatible with the reasons for the initial collection of the data using public powers. Hence, it seems that the Cabinet Office regards that the consent of the data subjects of the data for its further processing would not be required. Notably, Article 6(4) of the GDPR, in an attempt to identify when further processing would be compatible with initial processing, only mentions encryption and pseudonymization in its

[60] *Better use of data in government*, *supra* note 57, at 9.
[61] *Better use of data in government*, *supra* note 57, at 9.
[62] GDPR, *supra* note 1, at Article 4(5).
[63] *Better use of data in government*, *supra* note 57, at 3.

list of technical safeguards to be considered as relevant factors.[64] What is the UK Cabinet Office doing here? While it seems willing to operate within the framework of the data protection law, it does not rely on pseudonymization within the meaning of the GDPR as a compliance tool. Why? Well, probably because the definition to be found in the GDPR, as aforementioned, is not workable in practice.

Notably, the Digital Economy Bill[65] which followed the UK Cabinet Office consultation on data sharing was introduced in the House of Commons on 5 July 2016. Section 56 on Disclosure of information for research purposes provides that:

> (1) Information held by a public authority in connection with the authority‘s functions may be disclosed to another person for the purposes of research which is being or is to be carried out.
>
> (2) If the information is personal information it may not be disclosed under subsection (1) unless the following conditions are met.
>
> (3) The first condition is that, if the information identifies a particular person, it is processed before it is disclosed so that—
>
> (a)) the person‘s identity is not specified in the information, and
>
> (b) it is not reasonably likely that the person‘s identity will be deduced from the information (whether by itself or taken together with other information).[66]

While the language used in the Bill seems closer to that of pseudonymisation in the meaning of the GDPR, it is not entirely sure whether de-identified information is to be equated to data that has undergone pseudonymization.

F. THE ENISA REPORT ON BIG DATA

In its recent report on Privacy by Design and Big Data published in December 2015,[67] ENISA engages in the same type of analysis as Art. 29 WP, but without attempting to match legal definitions with technical definitions.

[64] To defend the position of the UK Cabinet Office, one could try to argue that this list is not exhaustive.

[65] Digital Economy Bill 2016-17, HC Bill [45] (U.K.).

[66] *Id.* at § 56.

[67] GIUSEPPE D‘ ACQUISTO ET AL., PRIVACY BY DESIGN IN BIG DATA: AN OVERVIEW OF PRIVACY ENHANCING TECHNOLOGIES IN THE ERA OF BIG DATA ANALYTICS (Eur. Union Agency For Network and Info. Security ed., 2015).

For background, ENISA was set up in 2004 with the goal of ensuring a high level of network and information security across the EU.[68] It acts as a centre of expertise in this area for the EU and works with its Member States, the private sector, as well as European citizens, to develop advice and recommendations on good practice in information security. In particular, it assists Member States in implementing relevant EU legislation and supporting the development of cross-border communities of expertise to this end.

In this report, ENISA takes the position that the concept of privacy by design is key to meet the challenges of technology for big data and meeting legal obligations.[69] It therefore explores anonymization in big data, together with other techniques and data access controls, which it deems are essential for providing data subjects with empowerment and control. ENISA starts by defining ―anonymisation‖ as referring to ―the process of modifying personal data in such a way that individuals cannot be re-identified and no information about them can be learned,‖[70] and describes the concept as stronger than that of de-identification ―which refers only to the removal of possible identifiers from the data set.‖[71]

On the other hand, however, it goes on in the report to distinguish what it calls, ―perfect anonymization‖ from ―low level anonymization,‖ which is ―usually not enough to ensure non-identifiability.‖[72] The consequence is that data that have undergone anonymization technique processes are given a very broad definition, so that it would seem they are not necessarily outside the scope of data protection laws. The report also introduces the concept of ―too strong anonymization‖ that ―may prevent linking data on the same individual (or on similar individuals) that come from different sources and, thus, thwart many of the potential benefits of big data.‖[73] Anonymization through aggregation might be an example of ―too strong anonymization.‖ Anonymization through pseudonymization would then be an alternative,

[68] *See generally* Regulation (EU) No 526/2013 of the European Parliament and of the Council of 21 May 2013 concerning the European Union Agency for Network and Information Security (ENISA) and repealing Regulation (EC) No 460/2004 (repealing Regulation 460/2004) O.J. (L 165), 18.6.2013, p. 41–58 (EU).

[69] PRIVACY BY DESIGN IN BIG DATA, *supra* note 67, at 22.

[70] PRIVACY BY DESIGN IN BIG DATA, *supra* note 67, at 27.

[71] PRIVACY BY DESIGN IN BIG DATA, *supra* note 67, at 27.

[72] PRIVACY BY DESIGN IN BIG DATA, *supra* note 67, at 27.

[73] PRIVACY BY DESIGN IN BIG DATA, *supra* note 67, at 27.

although the report does not refer to pseudonymization as such, but uses the concepts of pseudonyms and pseudonymity without putting forward workable definitions.[74]

In itself the report is therefore not that helpful to fully understand the breadth of legal definitions. In addition, it seems to be essentially focused on the data and the technology, without reference to other elements of the data environment or to any timeframe.

II. THE COMPONENTS OF A DYNAMIC APPROACH TO ANONYMIZATION

Probably the most influential legal piece on anonymization and its legal effects is Ohm's piece entitled ―Broken Promises Of Privacy: Responding to the Surprising Failure of Anonymisation,‖ which insisted in 2009 that de-identification is a failure and should be abandoned as a regulatory objective.[75] Ohm's highly-influential article treats what he calls ―release-and-forget anonymization‖ as an empty promise because he believes the relevant computer science literature proves the theoretical limits of the power of de-identification techniques.[76] Ohm argues that this requires replacing the underlying assumption of many privacy laws–-that anonymization is a ―silver bullet‖ to privacy problems—to ones premised on a more realistic assessment of the risks of re-identification and appropriate responses thereto. To this end, he also recommends abandoning the traditional distinction made between personal data, and non-personal (anonymized) data, in every privacy law and regulation that currently depends on them.

We reject this approach for two reasons. The first one is not a normative reason but a purely descriptive one; such an approach is not compatible with the GDPR as it still relies on the category of personal data to delineate its scope and does not exactly provide a plurality of regimes depending upon the risks of re-identification. The second reason is normative. Research is showing that if we agree that zero risk is not attainable, a comprehensive and ongoing assessment of data environments should still allow the implementation of robust

[74] PRIVACY BY DESIGN IN BIG DATA, *supra* note 67, at 22. Although ENISA does emphasize rightly the importance of linkability as a concept which privacy models must take in account when considering big data anonymization, PRIVACY BY DESIGN IN BIG DATA, *supra* note 67, at 32.

[75] Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701, 1744 (2010).

[76] *Id.* at 1755.

anonymization practices in satisfaction of an adequate level of legal protection of individuals‘ privacy and other rights that could be compromised when data relating to them are processed. This is not to say that perfect solutions have been found yet, but the effort seems promising.[77]

In addition, and we are echoing the findings of the UK Anonymisation Network which favours a ―clean separation between the complexities of data protection;‖[78] excluding a certain number of recipients from the category of data controllers, in particular researchers, is a way to simplify the regime. In particular, it would make the regime more easily understandable by private actors and especially data analysists and data scientists operating in the field given the intricacies of the law, e.g. in relation to data subject rights.[79] Moreover, excluding a certain number of recipients from the category of data controllers is likely to be more compliant with the data minimization principle itself: data controllers releasing datasets should be obliged to anonymize the data beforehand rather than recipients of the dataset such as researchers, which are actually required to pseudonymize to the extent possible, if not to anonymize under Article 89 of the GDPR. Furthermore, excluding a certain number of recipients from the category of data controllers would facilitate the transfer to researchers, who would still be required to comply with the framework established by the initial data controller, and would give an incentive to the latter to enter into a contractual relationship with recipients in order to mitigate the consequences of remaining a data controller.

The purpose of this section is thus to unfold the main components of a dynamic approach to anonymization which merges elements of a harm-based approach, a risk-based approach, and a procedure-based approach, and relies upon an examination of the data in context and over time. The point is made that such an approach is compatible with the newly adopted EU data protection framework, i.e. the GDPR.

A. COMBINING HARM-BASED, RISK-BASED, AND PROCEDURE-BASED

[77] *See generally* ELLIOT ET AL., *supra* note 6, at 67–119.

[78] ELLIOT ET AL., *supra* note 6, at 20.

[79] *See* GDPR, *supra* note 1, at chapter 3. While Article 14 of the GDPR contains an exception to the right to information in its paragraph 5, Article 15 of the GDPR does not and one has to go back to Article 11 to fully understand the contours of the right to access. *Id.*

APPROACH TOGETHER

Three types of approaches to the concept of anonymous data as delineated under data protection laws are emerging from the legal literature in the field, as described by Rubinstein and Hartzog:[80]

> ***Harm-based approach:*** This approach focuses on assessment of specific privacy harms that follow instances of poorly de-identified data, in particular relying upon the explicit detection of harm related to insufficient anonymization. Notwithstanding, successful re-identification attempts can be very difficult to establish ex post, not least because they often remain hidden. The approach also encounters difficulties as it depends upon establishing the causation of legally cognizable harm back to the poor anonymization process. The authors reject the harm-based approach for these reasons, not least because –harm is a contentious concept in privacy law‖; and, in practice, –many privacy harms are incremental or difficult to quantify and articulate‖ and –might not come to light until many years after the fact, if ever.‖[81]
>
> ***Traditional risk-based approaches:*** Under a risk-based approach, scholars argue that the definition of anonymous data in law should also be based upon a case-by-case assessment of the facts, but using an ex ante evaluation of the potential risks of re-identification on the basis of any given data in the particular circumstances.[82] Such an approach, moreover, implies that regular assessments of the risks associated with re-identification should be carried out, including when the data is reused. However, traditional risk-based approaches remain essentially output-based, which explains why the debate between –formalists and pragmatists‖[83] on de-identification techniques has not led to the wide-spread adoption of satisfactory data release policies.

[80] Ira S. Rubinstein & Woodrow Hartzog, *Anonymization and Risk*, 91 WASH. L. REV. 703 (2016).

[81] *Id.* at 730. The authors attribute the main source of this approach to –harm-based privacy regimes with high injury thresholds.‖ *Id.* at 730. In turn, they cite authors such as Ryan Calo in support of their rebuttal of this approach. Ryan Calo, *The Boundaries of Privacy Harm*, 86 IND. L. REV. 1131 (2011).

[82] As mentioned, the most notable advocate of such a risk-based approach is Paul Ohm. *See* Ohm, *supra* note 75, at 1751. Other notable proponents of this approach include Paul M. Schwartz & Daniel J. Solove. *See, e.g.*, Paul M. Schwartz & Daniel J. Solove, *The PII Problem: Privacy and a New Concept of Personally Identifiable Information*, 86 N.Y.U. L. REV. 1814, at 1887, 1894 (2011); Paul M. Schwartz & Daniel J. Solove, *Reconciling Personal Information in the United States and European Union*, 102 CAL. L. REV. 877, at 907, 909, 915 (2014); Omer Tene, *The Complexities of Defining Personal Data: Anonymization*, 8 DATA PROT. LAW & POLICY 8, 6 (2011); *and* W. K. Hon et al., *The Problem of 'Personal Data' in Cloud Computing: What Information is Regulated?—The Cloud of Unknowing*, 1 INT'L. DATA PRIVACY L. 211, 224 (2011).

[83] Rubinstein & Hartzog, *supra* note 80, at 714–17.

> ***Procedure-based approach****:* Instead of focusing on the ultimate goal of anonymization, under a procedure-based security approach, it is proposed that the law could be designed around processes necessary to lower the risk of re-identification and sensitive attribute disclosure. In other words, in common with other process-based regimes, the legal model mandates procedures, not outputs.[84] While aspects of the risk-based approach can be found in this model (for example, it requires a risk-assessment to be carried out prior to sharing data) or better it is still centered on the concept of minimising risks, it encompasses a risk-tolerant approach to data release policies, building on statistical disclosure limitation techniques.[85] This approach is risk-tolerant because the yardstick against which the acceptability of de-identification efforts under this model is assessed is the achievement of a level of reasonableness in adherence to industry standards.[86]

Borrowing from the field of data security, a procedure-based approach is preferred by Rubinstein and Hartzog because they consider it a sustainable strategy to focus on the preconditions and processes necessary for protection, not outputs. This is because, inter alia, this approach is focused on those who release data on behalf of its data subjects and their responsibilities to engage adequate procedures in that process. In this way, the authors believe that this approach will help move beyond the debate over the perfection (or lack thereof) of anonymization, towards an approach that ―utilizes the full spectrum of SDL [statistical disclosure limitation] techniques‖ as mechanisms of data control ―in conjunction with data treatment techniques, organizational support, and mindful framing to establish a sound deidentification regime.‖[87]

While it is true that evaluating and quantifying re-identification risks, just as with identifying and quantifying privacy harms, is not an easy task, a procedure-based approach should be informed by the results of an upfront evaluation of the likelihood and magnitude of such risks as might flow from the use of relevant data. To this end, the approach adopted by the ICO in its Code can be described as an early attempt at both a risk-based approach and a procedural approach in the above senses of these terms.[88] That is, a re-identification risk analysis can lead

[84] Rubinstein & Hartzog, *supra* note 80, at 733-34.

[85] Rubinstein & Hartzog, *supra* note 80, at 717-39.

[86] Rubinstein & Hartzog, *supra* note 80, at 736-37.

[87] Rubinstein & Hartzog, *supra* note 80, at 739.

[88] This is not to say that the ICO is not without flaws. While it has often been one of the first movers in many fields, such as anonymization, much still needs to be done in regards to

to the application of a set of protective measures accompanying the release of data, which is calibrated to the results of that analysis and deemed suitable to the specific data environment under consideration. The higher the risks, the stricter the procedural measures that should be applied, and vice versa.

Such an approach is not necessarily at odds with the ~~risk-based~~ approach advocated by Art. 29 WP. Excluding anonymized data from the scope of data protection laws when, and only when, certain safeguards are in place is not only compatible with making sure data subject rights are robust and systematically protected, but also with ensuring that fundamental principles remain applicable to controllers (i.e. legitimacy, data minimization, purpose limitation, transparency, data integrity, data accuracy). In its statement on the role of a risk-based approach in data protection legal frameworks,[89] Art. 29 WP was reacting against an approach advocating a change of focus of the regulation that should only be concerned with data use rather than with data collection. To be sure, applying a risk-based approach to the very definition of anonymized data is not tantamount to saying that only data use matters.

While Rubinstein and Hartzog do not formulate it exactly in these terms, a procedure-based approach would thus require a case-by-case approach and the construction of data environment models as a key step to identify the control measures to put in place to ensure that data is sufficiently protected such that it can justifiably be deemed to fall (for the time being) outside data protection laws. Mackey and Elliot define data environment in the following way, ―the data environment is made up of a small number of components: data, agents, and infrastructure. It is these components that we need to look at in order to ascertain how a statistical disclosure might occur and play out.‖[90]

enforcement. See, for example, criticisms about the apparent lack of enforcement conducted by the ICO after its decision to drop an investigation into a British Telecommunications plc data breach in 2011. *See* Derek du Preez, *ICO Under Fire for Dropping BT Data Breach Probe*, COMPUTING (Feb. 2, 2011), www.computing.co.uk/ctg/news/2023566/ico-dropping-bt-breach-probe). The ICO was also criticized for its freedom of information enforcement policy. *See* Matt Burgess, *FOI enforcement: The Four Enforcement Notices Issued in 10 years*, FOI DIRECTORY (July 2, 2015), www.foi.directory/featured/foi-enforcement-the-four-enforcement-notices-issued-in-10-years/.

[89] *See* Article 29 Data Protection Working Party, *Statement on the Role of a Risk-Based Approach in Data Protection Legal Frameworks* (European Comm'n, Working Paper No. 218, 14/EN, 2014) [hereinafter *Statement on the Role of a Risk-BasedApproach*].

[90] Mark Elliot & Elaine Mackey, *Understanding the Data Environment*, XRDS: CROSSROADS, THE ACM MAG. FOR STUDENTS, Fall 2013, at 36, 38.

According to the UKAN Decision-making Framework, anonymization is therefore not about removing all risk, but managing risk through taking precautions by carefully analyzing the data environment. It can broadly be divided into three main parts for consideration in sequential order when data is to be subject to anonymization techniques: the data context audit, the risk control and analysis, and the impact management.[91]

While it is implicit in the emerging literature on data environment,[92] the adoption of a holistic approach based on data environment models implies that the approach has to be dynamic. Such an approach is compatible with the newly adopted EU data protection regime.

B. EXAMINING THE DATA IN CONTEXT AND OVER TIME

The DPD as well as the GDPR are based on a contextual definition of personal data. This is clearly demonstrated by Art. 29 WP and decisions of the CJEU. What has not yet been clearly highlighted is whether the DPD and the GDPR can accommodate a dynamic approach to personal data and thereby to anonymized data. This holds true as long as one accepts that a risk-based approach is built within both the DPD and the GDPR. In this sense, as much as anonymization is not a property of the data taken in isolation,[93] personalization—meaning the characterization of data as personal data—should not be seen as a property of the data but as a property of the environment of the data,

[91] *See* ELLIOT ET AL., *supra* note 6, chapter 3 for a detailed analysis.

[92] *See e.g.*, Elliot, *supra* note 90; Raymond Heatherly et al., *Process-Driven Data Privacy*, Proceedings of The Twenty-fourth ACM International Conference on Information and Knowledge Management 1021–30 (2015); Catherine Heeney et al., *Assessing the Privacy Risks of Data Sharing in Genomics*, 14 PUB. HEALTH GENOMICS 17, 17–25 (2011); Isabelle Budin-Ljøsne et al., *DataSHIELD: An Ethically Robust Solution to Multiple-Site Individual-Level Data Analysis*, 18 PUB. HEALTH GENOMICS, 87–96 (2014); Chris Argenta et al., *Sensemaking in Big Data Environments,* Proceedings of the 2014 Workshop on Human Centered Big Data 53–55 (2014); Khaled El Emam et al., *Anonymising and Sharing Individual Patient Data*, BMJ (Mar. 20, 2015), http://www.bmj.com/content/350/bmj.h1139; Alex Endert et al., *Modeling in Big Data Environments*, Proceedings of the 2014 Workshop on Human Centered Big Data 56 (2014).

[93] The first key principle upon which the Anonymisation Decision-Making Framework is based is the following: ―You cannot decide whether data are safe to share/release or not by looking at the data alone.‖ ELLIOT ET AL., *supra* note 6, at 4.

particularly, when one moves away from the category of direct identifiers such as names or unique identifiers. [94]

To fully understand the implications of such a dynamic approach and the extent to which it can be said to be concordant with the new GDPR, it is crucial to revisit the very concept of personal data as it is defined under EU law. Despite its broadness, the category of personal data has some limits, explained by Art. 29 WP itself in its opinion on personal data[95] and the CJEU in two judgements. In other words, the definition of personal data is context-dependent.

As mentioned, Article 2(a) of the DPD defines personal data as ―any information relating to an identified or identifiable natural person (‗data subject')‖[96] specifying that ―an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity.‖[97] The CJEU confirmed the breadth of the category of personal data in its decision, Bodil Lindqvist.[98]

While the GDPR adopts a slightly different formulation, the category of personal data remains broad, if not broader. Article 4(1) reads as follows:

> ‗[P]ersonal data' means any information relating to an identified or identifiable natural person ‗data subject'; an identifiable 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, online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that person.[99]

With that said, while identifiability is a key component of the concept of personal data, it is crucial to understand that it is not the only one.

[94] An argument can even been made that as not all names should be considered personal data by content as explained below.

[95] 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).

[96] DPD, s*upra* note 1, at Article 2(a).

[97] DPD, *supra* note 1, at Article 2(a).

[98] Case C-101/01, Bodil Lindqvist, 6 Novembre 2003, ECLI:EU:C:2003:596 ¶ 27 (EU). *See also* Joined Cases C-92/09 & C-93/09, Volker und Markus Schecke GbR, Harmut Eifert v. Land Hessen, 9 Novembre 2010, ECLI:EU:C:2010:662 ¶ 52 (EU); Joined Cases C-468/10 & C-469/10, ASNEF, FECEMD v. Administracion del Estado, 24 Novembre 2011, ECLI:EU:C:2011:777 ¶ 42 (EU).

[99] GDPR, *supra* note 1, at Article 4(1).

Focusing on the ―relate to‖ component of the legal definition of personal data, which does not necessarily seem to be satisfied when the ―identifiability‖ component is satisfied, it becomes clearer that while the category of personal data is very broad, it is not all encompassing and context is actually essential. In its opinion on the concept of personal data of 2007,[100] Art. 29 WP breaks down the concept of personal data into four components (―any information‖; ―relating to‖; ―an identified or identifiable‖; ―natural person‖) and puts forward a three-prong test to determine whether the data at stake relates to a natural person. ―[I]n order to consider that the data ―relate‖ to an individual, a ―content‖ element OR a ―purpose‖ element OR a ―result‖ element should be present.‖[101] While these elements can be present at the same time, they are alternative criteria.

A content element is present when on the face of the data it is possible to identify a natural person. By way of example, a folder under the name of an individual containing his/her medical information is personal data by content. It would seem logical to include within this category not only names but also personal identifiers such as passport numbers. The answer might seem less obvious for quasi-identifiers or indirect-identifiers comprising attributes which taken independently do not allow identification (e.g. job title), but can allow identification when combined with other attributes (e.g. job title + birthday + car brand). While strictly speaking, these attributes would not be personal data by content they would very easily become personal data by purpose.

Despite the absence of a content element, a purpose element is present, when the purpose for processing the data is ―to evaluate, treat in a certain way or influence the status or behavior of an individual.‖[102] Importantly, the accuracy of the evaluation should be irrelevant. In this sense, so-called pseudonymized browsing information, such as the one collected in the UK Vidal case,[103] even if it is not personal data by content, is likely to be personal data by purpose.[104]

The last category of personal data is the most difficult to grasp and potentially the broadest. To fully understand its meaning, it is

[100] *Opinion 04/2007 on the Concept of Personal Data*, *supra* note 95, at 6.

[101] *Opinion 04/2007 on the Concept of Personal Data*, *supra* note 95, at 10.

[102] *Opinion 04/2007 on the Concept of Personal Data*, *supra* note 95, at 10.

[103] *See* Google Inc. v. Vidal-Hall [2015] EWCA (Civ) 311 (U.K.) (appendix to the judgement).

[104] *Opinion 04/2007 on the Concept of Personal Data*, *supra* note 95, at 11 (using a call log for a telephone as an example).

therefore necessary to strictly stick to the words of Art. 29 WP, which states:

> Despite the absence of a –content‖ or –purpose‖ element, data can be considered to –relate‖ to an individual because their use is likely to have an impact on a certain person‘s rights and interests, taking into account all the circumstances surrounding the precise case. It should be noted that it is not necessary that the potential result be a major impact. It is sufficient if the individual may be treated differently from other persons as a result of the processing of such data.[105]

To illustrate this third element, Art. 29 WP takes the example of monitoring information relating to taxis including location data and speed data. While the purpose of such processing is not necessarily to evaluate taxi drivers‘ conduct, in the words of Art. 29 WP it –can therefore have a considerable impact on these individuals, and as such the data may be considered to also relate to natural persons.‖[106] Why is it that such processing can have a considerable impact on taxi drivers? Most probably it is because of the nature of the information, the richness of the information—location data is combined with speed data—and the systematic nature of the processing. The further processing of this data, for a different purpose, is thus likely to have a considerable impact on these individuals. If one follows this logic, there might be an argument to sustain that so-called pseudonymized browsing information, depending upon its richness, and, even if it is not used to –to evaluate, treat in a certain way or influence the status or behavior of an individual‖[107] yet, is also personal data by result.

To take another example, one hot issue is whether IP addresses, by themselves, constitute personal data. IP addresses are numerical strings assigned to devices such as computers connected to a network relying upon the Internet Protocol for communication. The assignment can be done statically, one IP address is assigned to one device and remains the same each time the device connects to the network, or dynamically, for each connection the device is assigned an available IP address. As such, there is an argument that they are not personal data by content, while they can become personal data by purpose, in the same way as call log for telephones. Can they be considered in other cases personal data by result?

[105] *Opinion 04/2007 on the Concept of Personal Data*, *supra* note 95, at 11.

[106] *Opinion 04/2007 on the Concept of Personal Data*, *supra* note 95, at 11.

[107] *Opinion 04/2007 on the Concept of Personal Data*, *supra* note 95, at 6.

In its decision of 24 November, 2011 in the Scarlet case,[108] the CJEU stated that ―[t]hose [IP] addresses are protected personal data because they allow those users to be precisely identified.‖[109] While the CJEU does not really explain its finding, it is very likely that the CJEU was concerned about the combination of IP addresses, and content information relating to the types of copyright works accessed to or shared with by the subscribers of the Internet service provider, as well as ultimately subscriber information.[110] In this sense, one way of reading the CJEU‗s judgment is to say that the monitoring information including IP addresses meant to be collected by the ISP was personal data by result. Indeed, the further processing of this data for a different purpose and,
e.g. to determine whether subscribers are copyright infringers, would be likely to have a considerable impact on these individuals.

In order to determine whether IP addresses are personal data, it is thus essential to examine the nature of the information combined with these IP addresses as well as the purposes for which the data is processed. Said otherwise, a contextual assessment should be required in order to characterize IP addresses.

National courts, however, have not always been willing to engage into a detailed analysis of this sort to solve the issue of characterization and sometimes have artfully avoided this discussion altogether.[111] Surprisingly, Art. 29 WP itself has limited its analysis to broad statements without applying its three-prong test to this type of personal data. In its Opinion on the concept of personal data, it simply states under the heading of dynamic IP addresses that ―[t]he Working Party has considered IP addresses as data relating to an identifiable person,‖[112] referring back to a working document issued in 2000.[113]

[108] Case C-70/10, Scarlet Extended SA v. Société Belge des Auteurs, Compositeurs et Éditeurs S.C.R.L. (SABAM), ECLI:EU:C:2011:771 (Nov. 24, 2011)(EU).

[109] *Id.* at ¶ 51.

[110] Sophie Stalla-Bourdillon, *Online Monitoring, Filtering, Blocking . . . What is the Difference? Where to Draw the Line?*, 29 COMPUTER L. & SEC‗Y REV. 702, 708 (2013).

[111] Cour de cassation [Cass.] [supreme court for judicial matters] 1e civ., Nov. 3, 2016, Bull. civ. I, No. 1184 (Fr.) ; Cour d‗appel [CA] [regional court of appeal] Paris, 13e civ., Apr. 27, 2007.

[112] *Opinion 04/2007 on the Concept of Personal Data*, *supra* note 95, at 16.

[113] Article 29 Data Protection Working Party, *Privacy on the Internet – An Integrated EU Approach to On-line Data Protection* (European Comm‗n, Working Paper No. 37, 5063/00/EN, 2000), at 6. *See also* Article 29 Data Protection Working Party, *Opinion 1/2008 on Data Protection Issues Related to Search Engines* (European Comm‗n, Working Paper No. 148, 00737/EN, 2008).

The CJEU decided very recently in the Breyer case[114] to follow its Advocate General, who had considered that context was crucial to determine whether dynamic IP addresses should be considered as identifiable data and thereby characterized as personal data.[115] The CJEU referred to paragraph 68 of Advocate General Campos Sánchez-Bordona‘s opinion, which states:

> Just as recital 26 refers not to any means which may be used by the controller (in this case, the provider of services on the Internet), but only to those that it is likely ‗reasonably‘ to use, the legislature must also be understood as referring to ‗third parties‘ who, also in a reasonable manner, may be approached by a controller seeking to obtain additional data for the purpose of identification. This will not occur when contact with those third parties is, in fact, very costly in human and economic terms, or practically impossible or prohibited by law.[116]

The fact that the identifiability requirement, rather than the ―relate to‖ requirement, is used to characterize the data at stake in Breyer can only strengthen the claim that the characterization of personal data should be context-dependent.

Interestingly, Article 4(1) of the GDPR labels location data as an example of personal data without distinguishing between purposes.[117] Using the Art. 29 WP framework in its Opinion on the concept of personal data, it would seem that location data is either personal data by content or personal data by result. As location data is first of all data about an object, it might be that the second approach would make more sense. But isn‘t it the case that IP addresses are location data? Does it make sense to treat all types of location data in the same way?[118] Ideally,

[114] Case C-582/14, Patrick Breyer v. Bundesrepublik Deutschland, 19 October 2016, ECLI:EU:C:2016:779 (EU).

[115] Opinion of the CJEU Advocate General Campos Sánchez-Bordona, C‑ 582/14, Breyer v. Bundesrepublik Deutschland, 2016, at 68.

[116] *Id.* This statement should nevertheless be compared with the content of the following paragraph in which Advocate General Campos Sánchez-Bordona acknowledges that an Internet Service Providers is ―a main player in the structure of the Internet, who is known with certainty to be in possession of the data required by the service provider to identify a user.‖ *Id.* However, the CJEU does not refer to paragraph 69 of the Advocate General‘s opinion.

[117] GDPR, *supra* note 1, at Article 4(1).

[118] See, for example, the formulation of the version of the GDPR proposed by the European Parliament in its legislative resolution of 12 March 2014 (www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P7-TA-2014-0212+0+DOC+XML+V0//EN, retrieved on 17 May 2016), which had included IP addresses in its Recital 24 as an explicit example of an identifier whose processing should be subject to EU data protection rules (along with cookies and Radio Frequency Identification tags). Nonetheless,

the granularity of the location data should matter to determine whether it should be considered personal data by result. Such a granular approach does not seem to be necessarily precluded by the GDPR, even if the GDPR expressly refers to location data for the purposes of defining personal data.

In its YS decision, the CJEU states that a legal analysis is not personal data within the meaning of Article 2(a) of the DPD.[119] To reach this conclusion, the CJEU says two things, the second one being better formulated than the first. It first writes that ―[a]s regards, on the other hand, the legal analysis in a minute, it must be stated that, although it may contain personal data, it does not in itself constitute such data within the meaning of Article 2(a) of Directive 95/46.‖[120] The CJEU then recaps by saying, ―the data relating to the applicant for a residence permit contained in the minute and, where relevant, the data in the legal analysis contained in the minute are ‗personal data' within the meaning of that provision, whereas, by contrast, that analysis cannot in itself be so classified.‖[121] The second sentence is to be preferred to the first one because it shows that the legal analysis attached to the personal data by content (name, data of birth, nationality, gender, ethnicity, religion and language) is not personal data because it does not relate to the data subject but is ―information about the assessment and application by the competent authority of that law to the applicant's situation.‖[122] Once again, context is crucial and identifiability is not the only criterion at stake to characterize personal data.

The foregoing shows one important thing. Personalization is contextual and ultimately the concept of personal data has some limits. By context, one must understand the data itself, its richness, but also the intention of the data controllers and present and possible future uses. To simplify, at a minimum, it is essential to have an understanding of the purposes and of data linkages (i.e. whether and how it is or will be combined with other types of data) to assess the status of the data at stake. As a result, the line between personal data and non-personal data is fluid and evolves over time. Whether the inclusion of the ―relate to‖

the text goes straight on to caveat this statement with the parenthesis, ―unless those identifiers do not relate to an identified or identifiable natural person‖.

[119] Joined Cases 141 & 372/12, YS v. Minister voor Immigratie, Integratie en Asiel v. M,S, 17 July 2014, ECLI:EU:C:2014:2081, ¶ 48 (EU).

[120] *Id.* at ¶ 39.

[121] *Id.* At ¶ 48.

[122] *Id.* at ¶ 40.

component, as illustrated above, was intended to open widely the doors of the category of personal data after the adoption of some more restrictive approaches at the national level[123] is a moot point left open by this paper, as well as whether both the –relate to‖ and the –identifiability‖ components would need to be reconceptualized to better reflect context.

In as much as the category of non-personal data is context-dependent, we argue the same should be true for the concept of anonymized data. Thus, this should mean that as much as non-personal data can become over time personal data because the context changes over time, anonymized data can become personal data again. Besides, such a fluid line between the categories of personal data and anonymized data should be seen as a way to mitigate the risk created by the exclusion of anonymized data from the scope of data protection law. As a result, the exclusion should never be considered definitive and should always depend upon context.

Going back to the concept of anonymized data, even if at time n, the dataset has satisfactorily been anonymized taking into account privacy harms, risks of re-identification, and whether an appropriate procedure has been followed to reduce the amount of data to be transferred, and limit the number of recipients, at time n+1, the purpose of the further processing might be either to re-identify the data subjects or –to evaluate, treat in a certain way or influence the status or behavior of an individual,‖[124] which we take as meaning singling out. If this is the case, the once-anonymized data necessarily becomes personal data again and the further processing is subject to data protection law. Besides, it is very likely that the further processing of personal data will be considered as incompatible with the initial processing, and that no legal basis will be available, except maybe if the ultimate goal is to test the strength of the anonymization practice implemented and safeguards are put in place such as stringent security measures, including both organizational and technical security measures.

In the same line, if at time n+2, the anonymized dataset is combined with other datasets so that the processing of the data would have a considerable impact on data subjects (i.e. if the data could then be easily used to determine whether data subjects are compliant with the law), the once-anonymized data becomes personal data again. As stated above, uncertainty may persist over whether the further processing of

[123] *See e.g.,* Durant v. Fin. Serv. Auth. [2004] FSR 28, [2003] EWCA Civ 1746 (U.K.).

[124] *Opinion 04/2007 on the Concept of Personal Data, supra* note 95, at 6.

personal data will be seen as compatible with the initial processing. If the further processing is incompatible with the initial processing it will not be an easy task to legitimize it in compliance with EU data protection laws, save on the ground of a new legal basis.

In consequence, both at times n+1 and n+2, the anonymized data become personal data again and we end up in a situation of joint controllership. Both the person responsible for the anonymization of the dataset and the recipient of the dataset responsible for combining the different sets and/or with the intention to identify or evaluate data

Under Article 82(4), each controller –shall be held liable for the entire damage in order to ensure effective compensation of the data subject.‖[125] However, while Article 26 of the GDPR caters for situations of joint controllership, it envisages one specific situation: a situation in which both controllers are able to determine in a transparent manner their respective responsibilities for compliance. Pushing the analysis one step further, the first data controller, who made the decision to share the anonymized data set, should enter into contractual agreements with the recipients of the dataset, so that the latter put in place the necessary security measures when prescribed by the initial data controller and so that they are fully aware that they could ~~become~~ eventually become data controllers. Even if –the data subject may exercise his or her rights under this Regulation in respect of and against each of the controllers,‖[126] the first controller is then therefore entitled to have an action against the second controller(s) as per Article 82(5) if the latter is responsible for the damage and vice versa.

What is less clear is whether the first data controller could be seen as bearing an ongoing duty to monitor the data environment of anonymized datasets. If we assume that to determine whether a dataset is anonymized the answer has to be contextual and because context evolves over time, it can only make sense to subject data controllers to an ongoing monitoring duty, even if the dataset is considered to be anonymized, as per definition, initial data controllers are still data controllers. To be clear, the finding of such a duty is not necessarily in contradiction to the GDPR. If the characterization of anonymized data propels the data outside the remit of data protection laws and the principles of protection shall not apply to data rendered anonymous in such a way

[125] GDPR, *supra* note 1, at Article 82(4).

[126] GDPR, *supra* note 1, at Article 26(3).

that the data subject is no longer identifiable,[127] in the hands of the initial data controller, the anonymized dataset remains personal data as long as he has not destroyed the raw dataset that was then transformed to produce the anonymized dataset.

The aforementioned monitoring duty should require the data controller to keep pace with the evolution of technologies and communicating with data recipients to the extent possible. Clearly, such a duty will have a cost for data controllers wanting or having to release datasets. But at the very least, not all recipients of such datasets will be considered as data controllers themselves.

The foregoing shows that to make open data sustainable, one should move away from the release-and-forget model[128] or better from the release-and-complete-freedom model. Recipients of anonymized datasets should not be ignorant of data protection rules even if they are not necessarily data controllers or not yet data controllers.

III. CONCLUSION

To conclude, we argued in this paper that both the DPD and the GDPR rely on a risk-based approach for the very definition of anonymized data. This shall be true despite the ambiguous stance taken by Art. 29 WP in its Anonymization Opinion. Such a stance makes sense for at least two reasons. The first reason: the only way to make anonymized data a reality is to acknowledge that zero risk is not achievable. The second reason: given the homogeneity of the data protection regime, drawing a line between personal data and anonymized data would give a clearer message to those operating in the field. This does not mean that obligations would never be scalable, as Art. 29 WP recognizes it when it says:

> Implementation of controllers‘ obligations through accountability tools and measures (e.g. impact assessment, data protection by design, data breach notification, security measures, certifications) can and should be varied according to the type of processing and the privacy risks for data subjects. There should be recognition that not every accountability obligation is necessary in every case – for example where processing is small-scale, simple and low-risk.[129]

[127] GDPR, *supra* note 1, at Recital 26.

[128] Rubinstein & Hartzog, *supra* note 80, at 733, 739.

[129] *Statement on the Role of a Risk-Based Approach in Data Protection Legal Frameworks*, *supra* note 89, at 3.

This paper further makes the point that excluding anonymized data from the scope of data protection law is less problematic than first anticipated as the line between anonymized data and personal data should always remain a fluid line—anonymized data can always become personal data again depending upon the evolution of the data environment. Said otherwise, a dynamic approach to anonymized data is warranted. In this sense, the opposition between anonymized data and personal data is less radical than commonly described.

Such a dynamic and thereby contextual approach to anonymized data is compatible with the new data protection regime to be found in the GDPR. Both the DPD and the GDPR rely upon a contextual definition of personal data. Besides, and maybe more importantly, both the DPD and the GDPR can accommodate a dual characterization of a dataset depending upon in which hands the dataset is, i.e. those of the initial data controller in its raw form, or those of a subsequent recipient in its transformed state, and whether the initial data controller has put in place technical and organizational measures for the seclusion of the initial raw dataset transformed into an anonymized dataset. Furthermore, both the DPD and the GDPR could also accommodate an ongoing duty imposed upon the initial data controller to monitor the data environment of the transformed dataset.

What is crucial is to get the description of the data environment right for each processing activity and the modelling of data environment is obviously not a low-cost activity. Besides, more research is necessary in the field to fully comprehend the variety of categories of processing and the interplay between the different components of data environments: the data, the infrastructure, and the agents.

Finally, the paper shows that assuming a risk-based approach to anonymized data is a better route. As a result, it becomes essential to clearly delineate the contour of pseudonymized data, as the definition of pseudonymization to be found in the GDPR is misleading in that it does not refer to the risks associated with the linkability of individualized data records within one or across multiple datasets.

The research was partly funded by the European Union‘s Horizon 2020 research and innovation programme under grant agreement No 700542 and by the UK-GCHQ Effective Data Anonymisation Techniques Research Project. This paper reflects only the authors‘ views; the Commission is not responsible for any use that may be made of the information it contains.