Anonymous data v. Personal data—A false debate: An EU perspective on anonymisation, pseudonymisation and personal data

Sophie Stalla-Bourdillon and Alison Knight

PART I. 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 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 personal data that will be shared for further processing into "anonymous information" to use an EU legal term. "Anonymous information" is outside the scope of EU data protection laws, 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 leaves those in charge of processing the data with a problem: how to ensure that anonymisation is conducted effectively on the data in their possession, while retaining its utility for potential future disclosure to, and further processing by, third parties?

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

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

Yet, the texts of both the existing EU Data Protection Directive[1] (DPD) and the new EU General Data Protection Regulation[2] (GDPR) are ambiguous.

This paper suggests that, although the concept of anonymisation 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 EU data protection regulators) 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 are slightly more complex. Essentially, 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 that recipients of the transformed dataset are thereafter free from any obligations or duties because it 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.[3] Moreover, the state of anonymized data should be comprehended dynamically: anonymized data can become personal data again, depending upon the purpose of the further

[1] Directive 1995/46, 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) 31 (EC).

[2] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

[3] M. Elliot, E. Mackey, K. O'Hara and C. Tudor, The Anonymisation Decision-Making Framework, 2016, University of Manchester: Manchester.

processing and future data linkages, implying that recipients of anonymised data have to behave responsibly.

The paper starts by examining recent approaches to anonymisation, highlighting their shortcomings. It then explains why a dynamic approach to anonymisation is both more appropriate and compatible with the DPD and the GDPR. Ultimately, we conclude that the opposition between so-called "anonymous information" and personal data in a legal sense is less radical than usually described.

PART II. THE SHORTCOMINGS OF RECENT APPROACHES TO ANONYMISATION

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 anonymisation 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"[4] (Anonymisation Opinion). The Anonymisation Opinion was released two years after the release of the Code of Practice on "Anonymisation: Managing Data Protection Risks" (the Code)[5] by the UK Information Commissioner's Office (ICO) and departs from the ICO's Code on a significant point, as will be explained.

A. The DPD

Although Article 2(a) DPD suggests a very wide scope to the legal definition of personal data, the non-binding, but highly persuasive interpretation of it in Recital 26 of the DPD, appears to limit this definition using a "means likely reasonably" standard. Going further, the DPD appears to adopt a risk-based approach to personal data definition, and, thereby, to the legal effects of

[4] Opinion 5/2014 of the Article 29 Working Party on "Anonymisation Techniques," 2014.

[5] Information Commissioner's Office, Code of Practice on Anonymisation: Managing Data Protection Risk, (2012).

anonymisation 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 the data protection law.[6] 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.[7]

B. The UK ICO's Code of Practice on Anonymisation: Managing Data Protection Risk

The Code suggests, through its analysis, that if organizations takes reasonable security and disclosure limitation steps regarding data that has been subject to anonymisation techniques, its subsequent processing should not necessarily be caught by the UK Data Protection Act.[8] (Ultimately, whether it is caught will depend on assessment of the means likely reasonably standard as applied to the relevant circumstances).

The Code also distinguishes anonymisation output of non-individualized data resulting from data-aggregating processes, from processes removing certain identifiers from person-specific data but leaving individual-level information (carrying higher but not insurmountable risks to effective anonymization) . The latter includes pseudonymisation, defined as "distinguishing individuals in a dataset by using a unique identifier which does not reveal their 'real world' identity."[9] Given this approximation by the ICO between using unique identifiers and the non-revelation of real world identities, it is not clear reading the Code what is required to transform pseudonymised data into anonymised data.

---

[6] Recital 26, GDPR.

[7] Ibid.

[8] The Code, at 13.

[9] Ibid, at 29.

C. Art. 29 WP's opinion on anonymisation techniques

The Anonymisation Opinion describes pseudonymisation as a process by which one attribute—typically a unique one—in a record is replaced for another, not as a method of anonymisation, but merely a useful security measure.[10] This approach to pseudonymisation by Art. 29 WP appears better than the ICO's definition.

The Anonymisation Opinion includes statements that suggest that Article 29 WP is sympathetic to a risk-based approach.[11] Its position remains problematic, however, because - while presenting technical issues and risks inherent to anonymisation - Art. 29 WP also suggests that an 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 anonymisation process a few paragraphs earlier[12]. Moreover, Art. 29 WP states that data that has passed through an anonymisation 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.[13]

By affirming such statements, and despite other statements in the same Opinion, Art. 29 WP rejects 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

[10] Anonymisation Opinion, at 20.

[11] E.g. ibid, at 3, 4, 11-12, 25.

[12] Ibid, at 5.

[13] Anonymisation Opinion, at 9. The ICO does not agree with Art. 29 WP on this point (see The Code, at 13), in line with UK case law, see Common Services Agency v Scottish Information Commissioner [2008] UKHL 47, at [27, 92].

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.

D. The GDPR

Recital 26 of the GDPR clarifies that under the new regime, data protection principles will continue not to apply to anonymised data.[14] The GDPR still adopts, at least in its recital, a risk-based approach to anonymisation, relying upon the test of the "means reasonably likely to be used" by the data controller and third parties.

This said, the GDPR goes beyond the DPD by introducing a new definition: "pseudonymisation".[15] This definition is both narrow and very broad. It is narrow in that it excludes processes that cannot ensure that the personal data is not attributed to an identifiable natural person and this should be welcome.

More problematically, however, the definition is also very broad. As there is no reference to data linkability as the inherent problem belying concern that individuals may yet be singled out from data transformed by "pseudonymisation," it could include data that has undergone aggregation practices to remove individual-level elements within it. This, we suggest, is concerning.

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

[14] Recital 26, GDPR.

[15] Article 4(5), GDPR.

been destroyed, a transformed dataset must only be considered pseudonymised and remain subject to EU data protection laws. The GDPR would thus be endorsing Art. 29 WP's approach to anonymisation, which as explained, is not fully consistent with a risk-based approach to anonymisation.

A more nuanced interpretation of this sentence, building on the approach adoptedin the Code, by contrast, would be to say that if anonymisation through pseudonymisation seems to fall short legally, there still remains a route to effective anonymisation 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 singled out.

PART III. THE JUSTIFICATIONS FOR ADOPTING A DYNAMIC APPROACH TO ANONYMISATION

Probably the most influential legal piece on anonymisation is Ohm's piece entitled "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymisation" [16] which treats what he calls "release-and-forget anonymisation" as an empty promise.[17] He recommends abandoning the traditional distinction made between personal data and non-personal (anonymised) data.[18]

We reject this approach. First, purely descriptively, such an approach is incompatible with the GDPR as it still relies on the category of personal data to delineate its scope and does not provide a plurality of regimes depending upon the risks of reidentification. Second, more normatively, research shows that if we agree that zero-risk is not attainable, a comprehensive and

[16] Paul Ohm, Broken Promises Of Privacy: Responding To The Surprising Failure Of Anonymisation. 57 UCLA LAW REVIEW 1701, (2010).

[17] Ibid, at 1755 & 1756.

[18] Ibid, at 1743-1744.

ongoing assessment of data environments should still allow the implementation of robust anonymisation practices in satisfaction of an adequate level of legal protection of individuals' privacy. While perfect solutions remain elusive, the effort seems promising.[19]

In addition, to echo the findings of the UK Anonymisation Network which favours a "clean separation between the complexities of data protection;" [20] excluding certain recipients types from the category of data controllers, e.g. researchers, would simplify the regime. In particular, the regime would be more easily understood by private actors (especially data analysts and data scientists operating in the field) given the legal intricacies, e.g. in relation to data subject rights.[21] Moreover, excluding certain 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 anonymise the data beforehand (rather than dataset recipients, such as researchers, who are actually required to pseudonymise to the extent possible, if not to anonymise under Article 89 of the GDPR). Furthermore, excluding certain recipients from the category of data controllers would facilitate transfer to researchers, who would still be required to comply with the framework established by initial data controllers, and would give the latter incentives to enter into contractual relationships with recipients in order to mitigate the consequences of remaining data controllers.

To fully understand the implications of a dynamic approach to anonymization and the extent to which it can be said to be concordant with the GDPR, we must revisit the very concept of personal data as defined under EU law. Despite its broadness, the category of personal data has

---

[19] See once again MARK ELLIOT ET AL., THE ANONYMISATION DECISION-MAKING FRAMEWORK (2016). More work is nevertheless needed, in particular in relation to data situation modelling.

[20] Ibid, at 20.

[21] By way of example, 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.

some limits, as explained by Art. 29 WP in its opinion on personal data[22] and the CJEU in its judgment of 2014 in the YS case.[23] Said otherwise, the definition of personal data is context-dependent.

Article 2(a) of the DPD defines personal data as "any information relating to an identified or identifiable natural person ('data subject')"[24] 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."[25] The CJEU confirmed the breadth of the category of personal data in its decision, Bodil Lindqvist.[26] While the GDPR adopts a slightly different formulation, the category of personal data remains very broad, if not broader.[27]

Nevertheless, while identifiability is a key component of the concept of personal data, it is not the only one. 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 the category of personal data is not all encompassing and context is actually crucial. In its opinion on the concept of personal data of 2007,[28] Art. 29 WP breaks down the concept of personal data into four components ("any information"; "relating to"; "an identified or

[22] Opinion 4/2007 of the Article 29 Working Party on "the concept of personal data", 2007 (Personal Data Opinion).

[23] Joined cases C-141/12 and C-372/12, YS v Minister voor Immigratie, Integratie en Asiel and Minister voor Immigratie, Integratie en Asiel v M and S, (2015) ECLI:EU:C:2014:2081 (YS).

[24] Article 2(a), DPD.

[25] Ibid.

[26] Case C-101/01, Bodil Lindqvist, (2003) ECLI:EU:C:2003:596, at [27]. See also Joined cases C-92/09 and C-93/09, Volker und Markus Schecke GbR and Hartmut Eifert v Land Hessen (2010) ECLI:EU:C:2010:662, at [52 et seq]; and Joined cases C-468/10 and C-469/10, ASNEF, (2011) ECLI:EU:C:2011:777, at [42 et seq].

[27] Article 4(1), GDPR.

[28] Personal Data Opinion, at 6.

identifiable"; "natural person") and puts forward a three-prong test to determine whether relevant data 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."[29]

In its judgment in the YS case, the CJEU rules that a legal analysis is not personal data within the meaning of Article 2(a) DPD.[30] To reach this conclusion, the CJEU states "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."[31] This statement 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."[32]

Going back to identifiability, interestingly, Advocate General Campos Sánchez-Bordona in the Breyer case[33] seems to consider that, indeed, context is crucial for identifying personal data, and in particular characterising IP addresses as personal data. And the CJEU in its recent judgment of 2016 expressly refers to paragraph 68 of the opinion and thereby also excludes identifiability "if the identification of the data subject was prohibited by law or practically impossible on account of the fact that it requires a disproportionate effort in terms of time, cost and man-power, so that the risk of identification appears in reality to be insignificant."[34]

---

[29] Ibid, at 10.

[30] YS, at [48].

[31] YS, at [48].

[32] YS, at [40].

[33] Opinion of the CJEU Advocate General Campos Sánchez-Bordona, C-582/14, Breyer v Bundesrepublik Deutschland, (2016) ECLI:EU:C:2016:339, at [68].

[34] Case C-582/14, Breyer v Bundesrepublik Deutschland, (2016) ECLI:EU:C:2016:779, at [46].

In as much as the category of non-personal data is context-dependent, we argue the same should be true for the anonymised data concept. Such a fluid line between the categories of personal data and anonymised data should be seen as a way to mitigate the risk created by the exclusion of anonymised data from the scope of data protection law. Consequently, the exclusion should never be considered definitive but should always depend upon context. Ultimately, a key deterrent against re-identification risk is the potential re-application of data protection laws themselves.

Less clear is whether the first data controller could be seen as bearing an ongoing duty to monitor the data environment of anonymised datasets. If we assume that to determine whether a dataset is anonymised the answer has to be contextual, and because context evolves over time, it can only make sense to subject data controllers to ongoing monitoring duties, even if the dataset is considered anonymised, as per definition initial data controllers are still data controllers. To be clear, the finding of such a duty does not necessarily contradict the GDPR.

The next question is, then, whether contractual obligations between initial data controllers and dataset recipients are also crucial to fully control data environments and ensure re-identification risks remains sufficiently remote. It seems that they do indeed become crucial in cases in which it is essential for recipients of datasets to put in place security measures.

A dynamic approach to anonymisation therefore means assessing the data environment in context and over time and implies duties and obligations for both data controllers releasing datasets and dataset recipients. This raises the question whether the ICO got it right in the case of Queen Mary University of London of 2016.[35]

[35] Queen Mary University of London v (1) The Information Commissioner and (2) Alem Matthees, EA/2015/0269. For a comment, see S. Stalla-Bourdillon & A. Knight, blogpost, 19 September 2016, https://peepbeep.wordpress.com/2016/09/19/the-first-tier-tribunal-and-the-anonymisation-of-clinical-trial-data-a-reasoned-expression-of-englishness-which-would-have-to-be-abandoned-with-the-gdpr/.

PART IV. CONCLUSION

To conclude, we argue that both the DPD and the GDPR rely on a risk-based approach for the very definition of anonymised data. This shall be true despite the ambiguous stance taken by Art. 29 WP in its Anonymisation Opinion. We further posit that excluding anonymised data from the scope of data protection law is less problematic than first anticipated, as the line between anonymised data and personal data should always remain fluid: —anonymised data can always become personal data again depending upon evolving data environments. Said otherwise, a dynamic approach to anonymised data is warranted.

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. 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.