The GDPR and unstructured data: is anonymization possible?

Emily M. Weitzenboeck*, Pierre Lison**, Malgorzata Cyndecka***, and Malcolm Langford***

Key Points

- Much of the legal and technical literature on data anonymization has focused on structured data such as tables. However, unstructured data such as text documents or images are far more common, and the legal requirements that must be fulfilled to properly anonymize such data formats remain unclear and underaddressed by the literature.
- In the absence of a definition of the term 'anonymous data' in the General Data Protection Regulation (GDPR), we examine its antithesis—personal data—and the identifiability test in Recital 26 GDPR to understand what conditions must be in place for the anonymization of unstructured data.
- This article examines the two contrasting approaches for determining identifiability that are prevalent today: (i) the risk-based approach and (ii) the strict approach in the Article 29 Working Party's Opinion on Anonymization Techniques (WP 216).
- Through two case studies, we illustrate the challenges encountered when trying to anonymize unstructured datasets. We show that, while the risk-based approach offers a more nuanced test consistent with the purposes of the GDPR, the strict approach of WP 216 makes anonymization of unstructured data virtually impossible as long as the original data continues to exist.
- The concluding section considers the policy implications of the strict approach and technological developments that assist identification, and proposes a way forward.

Introduction

Big data is often characterized by its four constitutive 'Vs': digital data is produced in increasingly larger amounts (Volume), at high speed (Velocity), with a broad range of data types (Variety), and with differing levels of quality (Veracity).[1] This article focuses on the third dimension—Variety—and more specifically on the prevalence of unstructured or semi-structured data (such as text documents, images, or recordings) in most public and private organizations. According to some industry estimates, around 80 per cent of the world's data is unstructured.[2]

A large part of this unstructured content is likely to include personal data, making any processing of such data within the European Union (EU) framework subject to the General Data Protection Regulation (GDPR),[3] provided the processing is wholly or partly by

*Emily M. Weitzenboeck, Associate Professor, Faculty of Social Sciences, OsloMet – Oslo Metropolitan University, Norway; Education Fellow, Centre for Experiential Legal Learning (CELL), Faculty of Law, University of Oslo, Norway
**Pierre Lison, Senior Research Scientist, Norwegian Computing Centre, Oslo, Norway; Project Leader, CLEANUP (Machine Learning for the Anonymisation of Unstructured Personal Data)
***Malgorzata Cyndecka, Associate Professor, Faculty of Law, University of Bergen, Norway; Affiliate, Centre for the Science of Learning and Technology (SLATE), University of Bergen, Norway
***Malcolm Langford, Professor, Faculty of Law, University of Oslo, Norway; Adjunct Professor, Faculty of Law, University of Bergen, Norway; Director, Centre for Experiential Legal Learning (CELL), Faculty of Law, University of Oslo, Norway

1 In Lee, 'Big Data: Dimensions, Evolution, Impacts, and Challenges' (2017) 60(3) Business Horizons 293–303.

2 Juliette Rizkallah, 'The Big (unstructured) Data Problem' *Forbes* (5 June 2017).

3 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27April 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), OJ 2016 L 119/1. In addition to the material scope, for the GDPR to apply, the processing must also fall within the territorial scope of the regulation, cf art 3 GDPR.

automated means or, in the case of manual processing, as long as the data processed form part of or are intended to form part of a filing system.[4] Manual processing of files or sets of files, as well as their cover pages, which are not structured according to specific criteria thus fall outside the material scope of the GDPR.[5] However, since much unstructured data today are processed by automated means or, in the case of manual processing, are likely to be held in a filing system, very little unstructured data will fall outside the material scope of the GDPR. This means that processing of such data will thus only be permitted if it is done in a lawful, fair, and transparent manner and in accordance with the data protection principles underlying the GDPR.[6]

Furthermore, any further use of personal data that is incompatible with the purpose for which it was originally collected is only allowed if the data subject consents or if permitted by statutory law.[7] Admittedly, no new consent or separate legal basis is required in respect of further processing for archiving in the public interest, scientific or historical research purposes, or statistical purposesthis is not considered incompatible with the original purpose, cf Article 5(1)(b) GDPR. As long as technical and organizational measures pursuant to Article 89(1) GDPR (eg pseudonymization) are put in place, personal data may be further processed for such purposes. Otherwise, for all other new purposes, secondary use requires consent or a separate statutory basis, cf Article 6(4) GDPR. It is therefore not surprising that technologists, businesses, and the public sector are seeking to anonymize unstructured data—manually and increasingly computationally—so that any processing thereof will thereby fall outside the scope of the GDPR.

Although the GDPR adopts a broad view of what should be considered personal data, much of the legal and technical literature on the topic of anonymization has focused on what is often called structured data such as tabular databases. Structured datasets are characterized by a precise format that must be explicitly defined in advance and is strictly enforced by the database system. For instance, a tabular database is expressed as a list of 'records', each record being associated to a fixed, predefined set of attributes (such as age or nationality) and each attribute being associated to a predefined range of possible values (such as a positive number for the age, or the name of an existing country for the nationality).

However, structured datasets are only the tip of the data iceberg, and many types of data do not fit into such predefined formats. In particular, 'text documents'[8] may convey personal information through various linguistic formulations that are difficult to predict in advance. 'Images' can also express personal information through a broad spectrum of visual signals—most obviously when the image includes human faces or other identifiable features, but also through more indirect signals, such as pictures of vehicles with visible license plates. Similarly, 'audio recordings' may reveal personal information through acoustic patterns including both the voice of the speakers and the linguistic content that those speakers express. Unstructured data also include 'videos' and other 'multimedia content' which combine the above data types.

The common denominator between those unstructured data types is the fact that they do not follow a fixed, predefined template. As we shall see, this characteristic has important implications regarding the (im)possibility of conducting anonymization of such unstructured datasets in such a manner that they will no longer be deemed to constitute personal data pursuant to the GDPR (henceforth referred to as 'GDPR-compliant anonymisation' in this article).

The question of how to anonymize unstructured data such as texts or images (and whether this operation is at all possible in view of the requirements in the GDPR) has far-reaching consequences. Indeed, virtually all public or private organizations need to process and store unstructured data of some kind (including emails, case-handling notes, reports, recordings, pictures of various kinds, etc.). This notably includes data held by healthcare institutions, as much of today's medical information is only available in the form of text records such as clinical notes. Access to unstructured data including personal information is also a vital part of many

4 See art 2(1) GDPR. A 'filing system' is 'any structured set of personal data which are accessible according to specific criteria, whether centralised, decentralised or dispersed on a functional or geographical basis', cf art 4(6) GDPR. As explained by the Court of Justice of the EU in Case C-25/17 (Jehovah Witness), the analogous requirement in art 2(c) of the Data Protection Directive (95/46/EU) that the set of personal data must be 'structured . . . according to specific criteria' is 'simply intended to enable personal data to be easily retrieved'. Apart from that requirement, the article 'does not lay down the practical means by which a filing system is to be structured or the form in which it is to be presented'. See Case C-25/17 *Tietosuojavaltuutettu intervening parties Jehovan todistajat — uskonnollinen yhdyskunta* [2018] ECLI:EU:C:2018:551, paras 57 and 58.

5 See Recital 15 GDPR.

6 See art 5 GDPR.

7 See art 5(1)(b) and art 6(4) GDPR.

8 In this article, the term 'text document' shall be interpreted in a broad sense to encompass any type of data that includes free-form textual content. This definition includes therefore both formal communication (technical reports, clinical notes, web pages) as well as more informal, user-generated content (such as emails, blog posts or social media messages).

scientific fields, including social sciences, law, psychology, medical research and the humanities. However, as far as we can surmise, the legal scholarship on anonymization, whether it concerns the GDPR or national privacy laws throughout the world—is not focused on the less visible part of the data iceberg—unstructured data. The use of unstructured data is occasionally named, but not analysed in depth.[9]

This article makes two claims. First, that a risk-based approach to anonymization provides the most defensible interpretation of the GDPR and provides some space for the use of unstructured data after an anonymization process. Second, that, if we are to follow what is perhaps the most well-known interpretation on what to consider anonymous data according to the GDPR, ie Article 29 Working Party's Opinion 05/2014 on Anonymisation Techniques ('WP 216'),[10] the anonymization of unstructured data is essentially impossible. This impossibility in this strict approach does not primarily stem from the difficulty of masking direct and indirect identifiers in unstructured data (although, as we shall see, this task is far from trivial). Rather, the main legal difficulty resides in how the requirement of *non-linkability* between the anonymized data set and the original data source is interpreted in WP 216. When applied to unstructured data, this requirement is virtually impossible to satisfy due to the presence of various patterns (such as the occurrences of specific words or phrases in text documents) that can be exploited to link back an anonymized dataset to its original source. Consequently, if one were to follow the dictates of WP 216, the only remaining solution to obtain GDPR-compliant anonymizations of unstructured data is to effectively delete the original dataset, a measure that is typically unfeasible for most data controllers and would in many cases contravene other legal provisions.[11] Thus, we conclude that if this strict approach is the preferred approach, then WP 216 needs to be revisited as part of a public policy process.

This article is structured as follows: The section 'Technical Definitions' gives a broad overview of the technical issues by first drawing a distinction between structured and unstructured data and then highlighting some of the technical challenges encountered with the anonymization of unstructured data. The section 'Anonymization and Identifiability' delves into the legal issues. It first examines the notion of anonymous data from the point of view of it being the antithesis of personal data. It subsequently analyses the two contrasting approaches for determining identifiability that are prevalent today, ie the risk-based approach and WP 216's zero-risk strict approach, after which there is a discussion of the relevant agents of identification vis-à-vis the identifiability test. The section 'Case Studies' presents two case studies to illustrate the challenges encountered in the process of anonymization of unstructured data, in particular if the rigid interpretation of WP 216 is to be followed. The section 'Discussion and way forward' discusses the legal challenges highlighted in the preceding sections 34, and proposes a way forward.

Technical definitions

Structured and unstructured data

A common distinction in the field of data science is between 'structured' and 'unstructured' data.[12] This distinction rests on how the data is formatted: while structured data depends on a predefined 'data model', unstructured data does not follow a specific, predefined template.

A data model is a precise specification of how data is to be encoded. The easiest and most common data model is probably the 'table' where each row corresponds to a given record, and each column to an

9 For example, Kshetri commented that, 'Most organizations lack mechanisms to ensure that employees and third- parties have appropriate access to unstructured data and they are in compliance with data protection regulations': Nic Kshetri, 'Big Data's Impact on Privacy, Security and Consumer Welfare' (2014) 38 Telecommunications Policy 1134, at 1138. After making a similar point, Cumbley and Church note in passing that the existing data protection laws might be too harsh for unstructured data: 'Therefore Big Data provides a useful focus for many of the issues currently facing the privacy community and might suggest the need for more, or at least, tighter regulation. However, each step of the Big Data lifecycle – collection, combination, analysis and use – is already regulated by a current privacy framework which addresses most concerns and provides a sensible balance between the risks and benefits of Big Data. In fact, the more compelling case is for less regulation, particularly in relation to unstructured electronic data, which is the predominant reason for the growth of Big Data.': Richard Cumbley and Peter Church, 'Is "Big Data" creepy?' (2013) 29(5) Computer Law & Security Review 601. In their technical paper, Francopoulo and Schaub examine technical difficulties with anonymizing text data and propose a pseudonymization technique to deidentify text data, but stop short from a legal analysis of anonymization: Gil Francopoulo and Léon-Paul Schaub, 'Anonymization for the GDPR in the Context of Citizen and Customer Relationship Management and NLP', Proceedings of the workshop on Legal and Ethical Issues (Legal 2020) 9-14, <https://hal.archives-ouvertes.fr/hal-02939437/document> accessed 10 September 2021.

10 Art 29 Working Party, 'Opinion 05/2014 on Anonymisation Techniques' (WP 216, 10 April 2014).

11 Healthcare institutions are for instance required to retain an archive of their patient records and cannot freely delete them. See, for example, section 7 of the Norwegian Regulations on the Norwegian health archives and the Health Archive Register (Regulations of 18 March 2018 No 268). The requirement to retain such records in an identifiable manner means that such retention remains within the scope of the GDPR.

12 See also Borko Furht and Flavio Villanustre, 'Introduction to Big Data' in Borko Furht and Flavio Villanustre (eds), *Big Data Technologies and Applications* (Springer Switzerland 2016) 3–11).

Table 1. Example of personal data expressed in a tabular format

	Person name	Date of birth	Gender	Nationality	Vaccination Status
1	Peter Higgs	30.07.1975	Male	British	2 shots
2	Andreas Sauner	02.10.1981	Male	German	No shot
3	Laurence Barrière	03.10.1957	Female	French	1st shot

attribute. As shown in Table 1, tabular data imposes several structural constraints to the records. The first constraint is that the attributes associated with each record must be fixed and defined in advance. In the example of Table 1, each record is associated with five attributes (name, date of birth, gender, nationality, and vaccination status). Furthermore, each attribute also has predefined constraints as to the type of values that are permissible. For instance, the date of birth of a living person must be a valid date between 1900 and 2021. Similarly, the nationality only takes a predefined range of possible values.

This predefined structure considerably facilitates the anonymization process, as it provides a clear, unambiguous specification of what is known about each individual. Attributes that correspond to direct identifiers (such as person names) must be systematically erased, while attributes such as date of birth, gender, and nationality are characterized, in the field of computer science, as 'quasi-identifiers'[13]—which means that they do not typically single out an individual when considered in isolation but may do so when combined with one another and linked with background knowledge.[14] Finally, since the attributes of a given table are all defined in advance, data controllers can easily determine which attribute should be considered as requiring additional protection—as in the last column of our example, which contains health information and belongs therefore to the special categories of personal data pursuant to Article 9(1) GDPR.

There exist other types of data models beyond tabular structures. Another important data model that is widely used in computer science is the 'graph' (sometimes called a 'network'), which is composed of a set of 'nodes' and 'edges' between those nodes.[15] For instance, a graph can be used to express relations between individuals on social media or between adjudicators in legal cases.[16] Although such graphs are typically more expressive than tabular databases, they are still required to follow certain structural constraints (for instance, an edge must always be defined between two nodes).

In contrast, 'unstructured data' is not bound by a specific, predefined data model. The most common type of unstructured data are text documents written in 'natural languages' such as English or Chinese.[17] Although texts are generally expected to follow certain linguistic and stylistic conventions (such as adhering to the syntax of the chosen language, or starting a document with a title), those are just social conventions, and a text document may in theory consist of any possible sequence of words or characters. Furthermore, in contrast to tabular databases and other types of structured data, the expressivity of natural languages makes it possible to express the same semantic content in multiple ways.

If we consider again the example from Table 1, personal information on the same three individuals may be expressed in text form in the following manner: Peter Higgs, born on July 30, 1975, is a UK national and has already received 2 shots of the vaccine, while his German colleague Andreas Sauner, who will celebrate his 40th birthday on October 2, did not yet receive any shot. Meanwhile, their common acquaintance Laurence Barrière recently got her first vaccine shot. Mrs. Barrière

13 See Josep Domingo-Ferrer, David Sánchez and Jordi Soria-Comas, *Database Anonymization: Privacy Models, Data Utility, and Microaggregation-based Inter-model Connections* (Synthesis Lectures on Information Security, Privacy & Trust, Morgan & Claypool Publishers California 2016).

14 For instance, the combination of gender, birth date and postal code can be exploited to identify between 63 and 87% of the US population, due to the public availability of US Census Data, as first shown by Latayana Sweeney in her landmark study on re-identification of census data. See Latayana Sweeney, 'Uniqueness of Simple Demographics in the U.S. Population' (2000) Carnegie Mellon University, Laboratory for International Data Privacy, and Philippe Golle, 'Revisiting the Uniqueness of Simple Demographics in the US Population' (2006) *Proceedings of the 5th ACM Workshop on Privacy in electronic society* 77–80.

15 See, eg Stanley Wasserman and Katherine Faust, *Social Network Analysis: Methods and. Applications* (CUP Cambridge 1994).

16 Malcolm Langford, Daniel Behn and Runar Lie, 'The Revolving Door in International Investment Arbitration' (2017) 20(2) JIEL 301.

17 The term 'natural languages' is typically used to distinguish those from programming or mathematical languages, which have a much stricter set of constraints.

is French and will turn 64 years old on October 3. Although the content of the short text above is virtually identical to Table 1, a large part of the data's internal structure (such as the name and values of each attribute) is now implicit. The text also illustrates the occurrence of linguistic variations, as the gender, age and nationality of the three individuals can be expressed, either explicitly or implicitly (as in the use of the pronouns 'his' and 'her', which reveal the gender of the person being referred to). It should also be noted that, while a structured database typically contains one record per individual, a text document may simultaneously express personal information about multiple individuals and their relations to another. Indeed, the text indicates something that the table does not, namely the three individuals know each other: two are colleagues and they are acquainted with the third.

Unstructured data are not restricted to text documents and encompass (among others) images and audio-visual recordings. As for texts, those types of data often have a rich informational content but are not associated to a fixed, predefined data model: an image may consist of any combination of pixels, and an audio recording can store any sequence of sound signals.

It is worth noting that, in their definition of what constitutes a 'dataset', WP 216 adopts a restrictive definition that only seems to embrace structured data types:

> This opinion uses the following vocabulary in this section: a dataset is composed of different records relating to individuals (the data subjects). Each record is related to one data subject and is composed of a set of values (or "entries", e.g.: 2013) for each attribute (e.g. year). A dataset is a collection of records that can be shaped alternatively as a table (or a set of tables) or as an annotated/weighted graph, which is increasingly the case today.[18]

This impression is bolstered by the fact that all the examples discussed in WP 216 are of techniques applied to structured datasets.[19] As the section 'Anonymization of unstructured data: main challenges' of this article shows, the anonymization of unstructured data presents other challenges. The relevance and utility of WP 216 for anonymization of unstructured data is thus questionable.

Anonymization of unstructured data: main challenges

Like structured data, unstructured data often includes personal information. Text, images, and recordings may mention various individuals through direct and indirect identifiers and may also provide a variety of sensitive attributes (such as health conditions) about those individuals. However, one important challenge to address when one wishes to remove personal information from text or speech is that natural language is inherently ambiguous. A given word or phrase may have a different meaning according to the context. For instance, 'Pierre' may refer to a person's first name (in which case it would constitute a personal identifier) but also corresponds to the French word for stone. Consequently, the anonymization of unstructured data needs to take contextual factors into consideration upon deciding which part of the data may contribute to the risk of disclosing personal information. Although various computational approaches based on machine learning techniques have been devised in the past decade to automatically detect direct and indirect identifiers from text,[20] this task remains a difficult technological problem, and there is no approach (whether automated or manual) able to guarantee that all identifiers have been duly masked.

Similar challenges arise upon processing images or videos. Although a range of technological solutions have been developed to detect and blur certain visual traits such as human faces,[21] the detection and masking of indirect identifiers is a much harder task. This is also the case for images that do not feature any individual. For instance, an image showing the private home of an individual may indirectly disclose their identity.

Unstructured data is often high-dimensional in nature. A text document can be expressed as a long sequence of words, and an image as a collection of pixels. Each word or pixel can therefore be seen as representing a particular 'dimension' in the (very large) space of possible documents or images. Such high

18 See WP 216 (n 10) 12.

19 The same definition of 'dataset' is adopted by the Norwegian Data Protection Authority in its guidance document on anonymisation. See Datatilsynet, 'Anonymisering av personopplysninger: Veileder', 2015, 15 <www.datatilsynet.no/globalassets/global/dokumenter-pdfer-skjema-ol/regelverk/veiledere/anonymisering-veileder-041115.pdf> accessed 20 May 2021.

20 See, among others: Franck Dernoncourt and others, 'De-identification of Patient Notes with Recurrent Neural Networks' (2017) 24(3) Journal of the American Medical Informatics Association 596; Malcolm Langford, Runar Lie and Daniel Behn, 'Stylometric Analysis and Machine Learning: The Case of Investment Treaty Arbitration' in Ryan Whalen (ed), *Computational Legal Studies* (Edward Elgar Cheltenham 2020) 53.

21 See eg Z Ren, YJ Lee and MS Ryoo, 'Learning to Anonymize Faces for Privacy Preserving Action Detection' in *Proceedings of the European Conference on Computer Vision (ECCV)* (2018) 620–36.

dimensionality is also present in many structured datasets, as is notably the case for geolocation data collected on mobile devices. However, although high-dimensional structured datasets are also known to be challenging to anonymize,[22] they are nevertheless tied to a data model that explicitly defines the attributes associated with each individual. For instance, geolocation data will often be represented as sequences of spatial coordinates coupled with timestamps and device identifiers. In contrast, the 'dimensions' associated with a document or image do not directly express attributes associated with an individual. Due to this implicit and ambiguous mapping between the dataset itself and the personal information it may convey (in many circumstances, we do not even know which individuals may be referred to in a particular document), the anonymization of unstructured data requires the use of dedicated techniques that are often markedly different from the ones employed for other types of high-dimensional data.

Structured and unstructured data also differ in the types of anonymization operations (such as data suppression, generalization, perturbation, or aggregation) that can be applied upon them. In particular, while structured data can be aggregated (for instance by reducing a dataset to a set of key statistics derived from it), this is rarely possible for unstructured data. The range of possible techniques that can be employed to reduce the risk of re-identification is thus substantially more limited for unstructured data, where masking techniques are often the only feasible option.

Finally, in addition to the challenges related to the removal of direct and indirect identifiers that may allow an attacker to 'single out' a given individual, the anonymization of unstructured data needs to address another challenge, namely the possibility to link back the 'anonymised' dataset to its original source. In the section 'Case studies', we demonstrate empirically how this linkage can be performed, based on two case studies that focus respectively on text documents and medical images.

Anonymization and identifiability

Anonymous data as the antithesis of 'Personal Data'

Although the GDPR defines 'personal data', it contains no definition of 'anonymous data'. During the legislative process of the GDPR before the European Parliament, the rapporteur's draft report introduced a definition of 'anonymous data' which explicitly excluded such data from the scope of the GDPR.[23] However, the proposed definition was later removed from the European Parliament's Committee on Civil Liberties, Justice and Home Affairs' (LIBE Committee) compromise text and was not included in the final text of the GDPR.[24] Though the term is not defined in the GDPR, in data protection discourse, data that is not personal data is typically referred to as anonymous data.[25] Anonymous data is the antithesis of personal data. Effective anonymization thus depends on a sound understanding of what constitutes personal data.[26]

The term 'personal data' is the cornerstone of data protection legislation. Only information that constitutes 'personal data' falls within the scope of the GDPR. Personal data is defined in article 4(1) of the GDPR as:

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

22 YA De Montjoye and others, 'Unique in the Crowd: The Privacy Bounds of Human Mobility' (2013) 3(1) Scientific Reports 1–5.

23 Amendment 14 proposed the following additional text to Recital 23 (now Recital 26): 'This Regulation should not apply to anonymous data, meaning any data that cannot be related, directly or indirectly, alone or in combination with associated data, to a natural person or where establishing such a relation would require a disproportionate amount of time, expense, and effort, taking into account the state of the art in technology at the time of the processing and the possibilities for development during the period for which the data will be processed'. See European Parliament, Committee on Civil Liberties, Justice and Home Affairs (LIBE), 'Draft report on the proposal for a regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation) 2012/0011(COD), Rapporteur: Jan Philipp Albrecht.

24 European Parliament, 'Draft report on the proposal for a regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation) (COM(2012)0011 – C7-0025/2012 – 2012/0011(COD)) Committee on Civil Liberties, Justice and Home Affairs, Rapporteur: Jan Philipp Albrecht.

25 Lee A Bygrave and Luca Tosoni, 'Article 4(1)' in Christopher Kuner, Lee A Bygrave and Christopher Docksey (eds), *The EU General Data Protection Regulation (GDPR): A Commentary* (OUP Oxford 2019) 105.

26 This was also affirmed by the UK Information Commissioner. See ICO, 'Anonymisation: Managing Data Protection Risk – Code of Practice' (2012) 11.

In its Opinion 04/2007 on the Concept of Personal Data ('WP 136'),[27] the Article 29 Working Party analysed the term 'personal data' by breaking it down into its four chief constitutive elements, an approach that has become commonplace in data protection literature: 'information' that 'relates' to an 'identified/identifiable' natural 'person'.[28] The constitutive elements of 'anonymous information' are thus the negation of the four chief constitutive elements of 'personal data'. Anonymous information is thus: (i) information which does not (ii) relate to (iii) an identified or identifiable (iv) natural person. This also reflects the description of 'anonymous information' in Recital 26 GDPR as 'information which does not relate to an identified or identifiable natural person'.

Pseudonymized data, ie personal data that has undergone a process of pseudonymization, is still attributable to a natural person and thus subject to the GDPR.[29] However, 'personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable' falls outside the GDPR, cf Recital 26 GDPR. As does data which at its origins 'did not relate to an identified or identifiable natural person, such as data on weather conditions generated by sensors installed on wind turbines or data on maintenance needs for industrial machines'.[30] Where, however, non-personal data such as, for example, weather data, is likely to be used to assess its influence on individual personal behaviour, identification is intended, the data becomes information relating to people in purpose, and is thus personal data.[31] This is not as farfetched as it may seem. One can envisage other situations where non-personal data such as data on precision farming which can help to monitor and optimize the use of pesticides and water, to take the example mentioned in Recital 9 of the Free Flow of Non-Personal Data Regulation,[32] may be linked to other data that identify the individual farmers who may have introduced such innovative techniques in small farms. The Free Flow of Non-Personal Data Regulation recognizes the existence of mixed datasets, that is 'a data set composed of both personal and non-personal data' and states that:

> [i]n the case of a data set composed of both personal and non-personal data, this Regulation applies to the non-personal data part of the data set. Where personal and non-personal data in a data set are inextricably linked, this Regulation shall not prejudice the application of Regulation (EU) 2016/679.[33]

As explained by the European Commission, this implies that:

- the Free Flow of Non-Personal Data Regulation applies to the non-personal data part of the dataset;
- the General Data Protection Regulation's free flow provision[34] applies to the personal data part of the dataset; and
- if the non-personal data part and the personal data parts are 'inextricably linked', the data protection rights and obligations stemming from the General Data Protection Regulation fully apply to the whole mixed dataset, also when personal data represent only a small part of the dataset.[35]

Neither the GDPR nor the Free Flow of Non-Personal Data Regulation define the concept of 'inextricably linked'. According to the European Commission, 'it can refer to a situation whereby a dataset contains personal data as well as non-personal data and separating the two would either be impossible or considered by the controller to be economically inefficient or not technically feasible'.[36]

27 Article 29 Working Party, 'Opinion 04/2007 on the Concept of Personal Data' (WP 136, 20 June 2007).

28 See Bygrave and Tosoni (n 25) 109. See also, eg Michèle Finck and Frank Pallas, 'They Who must not be Identified – Distinguishing Personal from Non-personal Data under the GDPR' (2020) 10 International Data Privacy Law 11; Nadezhda Purtova, 'The Law of Everything. Broad Concept of Personal Data and Future of EU Data Protection Law' (2018) 10 Innovation and Technology 40.

29 Pseudonymization is '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', cf art 4(5) GDPR.

30 See Commission, 'Guidance on the Regulation on a framework for the free flow of non-personal data in the European Union' (Communication) COM (2019) 250 final, 6.

31 Purtova has illustrated how even weather may be deemed to be personal data. She gives the example of the Dutch Stratumseind 2.0 smart city project, which is a living lab. Among the aims of the project is predicting, preventing and de-escalating deviant behaviour on Stratumseind, a street in Eindhoven, the Netherlands, among other things, by engaging the police or adapting the street lighting. Various types of data are gathered from multiple sensors, including video- and acoustic cameras, sound sensors, WiFi tracking and a weather station. As she explains, 'one could argue that "if the weather is going to be used to target and categorise me, I need protection against its potential to define me as dangerous or depressed", even if achieving this protection is difficult. I agree.' See Purtova (n 28) 57–59.

32 Regulation (EU) 2018/1807 the European Parliament and of the Council of 14 November 2018 on a framework for the free flow of non-personal data in the European Union, OJ 2018 L 303/59.

33 Art 2(2) Free Flow of Non-Personal Data Regulation (n 32).

34 See art 1(3) GDPR.

35 Commission, 'Guidance on the Regulation on a framework for the free flow of non-personal data in the European Union' (n 30) 9.

36 Ibid 10.

As the above discussion shows, the scope of the term 'personal data' is very wide. This was an intentional and 'deliberate approach chosen by the legislator',[37] an approach hailed by the European Commission as having 'the benefit of flexibility, allowing it to be applied to various situations and developments affecting fundamental rights, including those not foreseeable' at the time that the 1995 Data Protection Directive,[38] which has a definition of personal data that is essentially the same as that in the GDPR, was adopted.[39]

Determining identifiability: conflicting approaches

Key to the notion of personal data is that an individual is identified or identifiable. Recital 26 GDPR lays down the criteria to determine identifiability in an identifiability test:

> 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 and technological developments.

The test is one of reasonable likelihood of identification either by the controller or by another person using state of the art technology, a test that is very much in line with the risk-based approach in data protection law.[40] However, as mentioned in the introduction to this article, a stricter, zero-risk approach was put forward in WP 216. The following sections examine each of these two approaches in turn, after which is a discussion of which agents of identification are relevant when applying the identifiability test.

The risk-based approach

The test in Recital 26 GDPR is based on the risk of identification and takes into account 'all objective factors', some of which are exemplified in the recital. According to a literal interpretation of Recital 26, where there is a reasonable risk of identification, data ought to be deemed to be personal data and treated as such. This implies that where that risk is merely negligible, 'data can be treated as non-personal data, and this even though identification cannot be excluded with absolute certainty'.[41] Indeed, in WP 136, the Article 29 Working Party stated that a 'mere hypothetical possibility' to single out an individual is not enough to consider that person as identifiable.[42] As the Working Party explained, the criterion of 'all the means likely reasonably to be used' by the controller or any other person in Recital 26 of the Data Protection Directive (DPD), which phrase is replicated in near identical wording in Recital 26 GDPR,[43] necessitates that 'all the factors at stake' should be taken into account.[44] Among such factors are the following:

- The cost of conducting identification;
- The intended purpose of processing (the implication here being that 'where the purpose of the processing implies the identification of individuals, it can be assumed that the controller or any other person involved have or will have the means "likely reasonably to be used" to identify the data subject');[45]
- The risk of organizational dysfunctions, (eg breaches of confidentiality duties) and technical failures;
- The state-of-the-art in technology at the time of processing, and the possibilities for technological developments during the lifetime of the processing;
- The technical and organizational measures that are in place to prevent identification, ie put in place as 'a *condition* for the information precisely not to be considered personal data' thereby falling outside the scope of the DPD;[46]
- The amount of time required for identification.

The first five factors abovementioned were highlighted in WP 136. Two of those factors—cost and the use of state-of-the-art technology throughout the lifetime of the processing—as well as the sixth factor abovementioned, ie the time required for

37 Commission, 'A comprehensive approach on personal data protection in the European Union' (Communication) COM (2010) 609 final, 5.

38 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, OJ 1995 L281/31.

39 Ibid. On the essential similarity of definitions, see Bygrave and Tosoni (n 25) 108.

40 See, eg Article 29 Working Party, 'Statement of the WP29 on the role of a risk-based approach in data protection legal frameworks' (WP 218, 30 May 2014).

41 Finck and Pallas (n 28) 14.

42 WP 136 (n 27) 15.

43 In GDPR Recital 26, the words 'reasonably likely' were inverted to bring the text more in line with ordinary grammatical convention. See Bygrave and Tosoni (n 25) 109.

44 WP 136 (n 27) 15.

45 Ibid 16.

46 Ibid 17.

identification, are now specifically included in Recital 26 GDPR as examples of objective factors that must be taken into account when determining whether means are 'reasonably likely to be used' to identify the individual. The list in Recital 26 GDPR is not exhaustive and thus all the abovementioned factors must be considered.

One way of assessing re-identification risk is by carrying out what in the field of information security is known as penetration testing, ie by replicating what a plausible motivated intruder might do and the resources he/she might have, to execute a re-identification and/or disclosure attack on the data.[47] Both the UK Information Commissioner's Office (ICO) and the Norwegian Data Protection Authority, for example, recommend the performance of what they refer to as a 'motivated intruder' test.[48] The 'motivated intruder' is characterized as a reasonably competent person who has access to resources such as the internet, libraries, and all public documents, and who is willing to employ investigative techniques such as actively making enquiries to uncover information. The 'motivated intruder' is not assumed to have any specialist knowledge such as computer hacking skills, or to have access to specialist equipment or to resort to criminality to gain access to the data.[49]

Looking solely at the data being subjected to anonymization techniques is insufficient to determine the extent of the re-identification risk. Besides the factors related to the data itself, the 'environment' in which the data are to be shared and released must also be taken into account.[50] Criteria linked to what, in the field of statistical confidentiality, is known as the 'data environment', have been proposed by the UK Anonymisation Network ('UKAN') to help interpret the concepts of 'personal data' and 'anonymisation', in particular with regard to the identifiability test.[51] Though UKAN's influential Anonymisation Decision-Making Framework is focused on structured data,[52] it re-directs the focus of re-identification risk to the wider context of the data environment and is thus also relevant to address some of the challenges encountered when de-identifying unstructured data. The 'data environment' is:

> the set of (formal or informal) structures, processes, mechanisms and agents that either (i) interact with the derived dataset; (ii) control interactions with that data; or (iii) provide interpretable context for that data.[53]

A data environment is thus deemed to consist of four key elements: other data, data users, governance processes, and infrastructure. The first element considers other data available in the environment in which the derived dataset is placed; the second element models how data users might act and operate on/in the data environment; governance processes (eg data access controls, licensing arrangements, contracts) determine how the users' relationships with the data are managed; and the infrastructure element considers how the physical and software processes implement functional restrictions on the environment. Various proponents of the risk-based approach claim that de-identified data accessed in a controlled environment, that is, in a situation where there is a combination of data and environment controls, should be deemed to be functionally anonymized.[54] The defining proposition of functional anonymization, according to these scholars, is the following:

> Whether data is anonymous or not (and therefore personal or not) is a function of the relationship between that data and its environment.[55]

ADR UK (Administrative Data Research UK) and the proposed *Helseanalyseplattformen* (health analytics platform) in Norway are examples of controlled environments that provide access to de-identified or anonymized data for research and statistical purposes.[56] In the case of ADR UK, only accredited or approved researchers are

47 Mark Elliot and others, *The Anonymisation Decision-Making Framework* (UKAN Publication Manchester 2016) 51. Penetration testing is also typically recommended by national security agencies such as the UK National Cyber Security Centre and the Norwegian National Security Authority. See <http://www.ncsc.gov.uk/guidance/penetration-testing> and <http://www.nsm.no/regelverk-og-hjelp/rad-og-anbefalinger/grunnprinsipper-for-ikt-sikkerhet-2-0/oppdage/gjennomfor-inntrengnings tester/> both accessed 28 May 2021.

48 See Information Commissioner's Office, *Anonymisation: Managing data protection risk - Code of practice*, (2012) 22–23 <www.ico.org.uk/media/1061/anonymisation-code.pdf> accessed 28 May 2021, 22–23 and Datatilsynet, *Anonymisering av personopplysninger: Veileder* (2015) 11–12 <www.datatilsynet.no/globalassets/global/dokumenter-pdfer-skjema-ol/regelverk/veiledere/anonymisering-veileder-041115.pdf> accessed 20 May 2021. Computer scientists sometimes refer to the intruder alternatively as an 'adversary', ie as someone who is motivated to do something that the data administrator wishes not to happen. See P Ohm, 'Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization' (2010) 57 UCLA Law Review 1701, 1723.

49 Elliot and others (n 47) 51.

50 Ibid 16 *et seq.*

51 Ibid.

52 Ibid 12.

53 Mark Elliot and others, 'Functional Anonymisation: Personal Data and the Data Environment' (2018) 34 Computer Law & Security Review 204, 2013.

54 Miranda Mourby and others, 'Are "pseudonymised" Data Always Personal Data? Implications of the GDPR for Administrative Data Research in the UK' (2018) 34 Computer Law & Security Review 222, 232. See also Elliot and others, ibid.

55 Elliot and others (n 53) 213.

56 See the UK ADR's webpages <www.adruk.org/about-us/about-adr-uk/> and the Norwegian Directorate of eHealth's webpages <www.ehelse.no/

given access to the de-identified and unpublished data for use in pre-approved research projects in the public interest. As a general rule, the data can be accessed only at certain physical locations and safe rooms, although, where data owner agreements are in place, some data is available to researchers via secure remote access.[57] As regards the proposed health analytics platform in Norway, access to health data on the platform is planned to be provided in secure spaces. Researchers will only be able to extract analytical results but will otherwise be unable to take out or download other data from the platform.[58]

The strict approach

In its Opinion on Anonymization Techniques (WP 216) issued a mere 7 years after WP 136, the Article 29 Working Party took a stricter approach and interpreted Recital 26 DPD very narrowly.[59] Although WP 216 has not been expressly endorsed by the European Data Protection Board (EDPB),[60] such document remains relevant and influential, and the EDPB itself refers to it in recent documents.[61] In WP 216, the Working Party examined the main anonymization techniques, ie randomization and generalization, and recognized that a risk factor is inherent to anonymization.[62] In that regard, it identified three risks that are essential to anonymization:

> *Singling out*, which corresponds to the possibility to isolate some or all records which identify an individual in the dataset;
> *Linkability*, which is the ability to link, at least, two records concerning the same data subject or a group of data subjects (either in the same database or in two different databases). If an attacker can establish (e.g., by means of correlation analysis) that two records are assigned to a same group of individuals but cannot single out individuals in this group, the technique provides resistance against "singling out" but not against linkability;
> *Inference*, which is the possibility to deduce, with significant probability, the value of an attribute from the values of a set of other attributes.[63]

To determine the robustness of one's anonymization technique and ultimately whether anonymization has occurred, all three risk criteria abovementioned must be taken into account.[64] A solution against these three risks would, according to the Working Party, be robust against re-identification attempts.[65] Though it assessed the strengths and weaknesses of various anonymization techniques using the three criteria abovementioned as a yardstick, the Working Party then applied what has been termed 'a zero-risk test'[66] when it stated that 'anonymisation results from processing personal data in order to irreversibly prevent identification'.[67]

So absolute is the Working Party's approach in WP 216, that it equates anonymization with erasure of data. According to the Working Party, a close reading of the DPD's Recital 26,[68] as well as the requirement in Recital 26 and Article 6(1) of the e-Privacy Directive to erase or anonymize traffic data ('erased or made anonymous'), and the requirement in Article 9(1) of the e-Privacy Directive that certain location data may only be processed when such data 'are made anonymous' or with the data subject's consent, means that 'the outcome of anonymisation as a technique applied to personal data should be, in the current state of technology, as permanent as erasure, i.e. making it impossible to process personal data'.[69]

Building up to a crescendo, WP 216 then states that, as long as the original (identifiable) data set exists, any resultant dataset to which anonymization techniques have been applied is still considered to be personal data:

> Thus, it 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

programmer/helsedataprogrammet/helseanalyseplattformen> accessed 2 June 2021.

57 See UK ADR <www.adruk.org/our-data/our-data-services/#c4811> accessed 2 June 2021.

58 As a main rule, access will be given to de-identified data. However, as discussed in the preparatory works to the amendments to the Norwegian Health Register Act which set up the platform, in certain exceptional cases, access may be given to identifiable personal data. Of course, access to identifiable personal data falls squarely within the GDPR's material scope. See preparatory works to the Health Register Act, Prop 63 L (2019–2020) section 12.5.9.5, 103.

59 According to Recital 26 DPD, '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'. Unlike Recital 26 GDPR, Recital 26 DPD did not contain any list of objective factors.

60 See EDPB, Endorsement 1/2018, adopted 25 May 2018 <https://edpb.europa.eu/sites/default/files/files/news/endorsement_of_wp29_documents_en_0.pdf > accessed 10 September 2021.

61 See EDPB, 'EDPB Document on response to the request from the European Commission for clarifications on the consistent application of the GDPR, focusing on health research', para 46, adopted on 2 February 2021 <https://edpb.europa.eu/sites/default/files/files/file1/edpb_replyec_questionnaireresearch_final.pdf> accessed 10 September 2021.

62 See WP 216 (n 10) 7.

63 Ibid 11–12.

64 Ibid 3.

65 Ibid 12.

66 Finck and Pallas (n 28) 15.

67 WP 216 (n 10) 3. Our emphasis.

68 The 'data should be such as not to allow the data subject to be identified via "all", "likely" and "reasonable" means'. See WP 216 (n 10) 5.

69 Ibid 6.

> this dataset (for example after removal or masking of identifiable data), the resulting dataset is still personal data.[70]

As Ohm succinctly puts it, '[d]ata can be either useful or perfectly anonymous but never both'.[71] The trade-off for achieving anonymization pursuant to WP 216 is thus complete destruction of the original data. This is an interpretation of anonymization so extreme that, as Ohm puts it, 'no data administrator would ever use it' since one ends up with 'a complete wiped database with absolutely no information beyond the single field of information under study' such as, in the case of a health study, perhaps the diagnosis, for an education study the grade point averages, and for a labour study the salaries.[72] As shown in the section 'Case Studies' of this article, we were faced with a similar dilemma when we applied WP 216's strict yardstick in a process of anonymization of two different types of unstructured data, viz. text documents and medical images.

Relevant agents of identification

According to Recital 26 GDPR, the means reasonably likely to be used 'by the controller or by another person' to identify an individual must be taken into account to determine whether the person is identifiable. The question of who the relevant agents of identification are in a situation where one person has data which does not 'per se' identify an individual while the additional data needed to identify the person to whom that data relates is in the hands of another person arose in *Breyer*. In *Breyer*, certain data relating to visitors of websites operated by Federal German institutions, namely, the dynamic IP address, date and time of access of a website, were stored by such institutions to ensure the security and continued proper functioning of their websites.[73] However, those institutions did not have the additional data necessary to enable them to identify the website visitors. That additional data was in the hands of the internet service provider ('ISP') that had allocated the IP address to the website users.

Though the facts in *Breyer* are not identical to those underlying this article in that *Breyer* did not deal with de-identified but with partial data, the question that arose in *Breyer* is nonetheless relevant to the discussion in this article. Can data that do not directly identify individuals and that are collected by an entity be said to contain (or constitute) personal data, in a case where another entity has the additional data required to identify the individual? In *Breyer*, the Court of Justice of the EU ('CJEU') was asked to determine whether the dynamic IP address in the hands of the Federal German institutions operating the websites in question was data relating to an identifiable personal data.

Both Advocate General Campos Sánchez-Bordona and the CJEU, as well as the *Bundesgerichtshof*, the referring court in *Breyer*, referred to two opposing views debated by German scholars: an 'objective' or absolute criterion and a 'relative' or subjective criterion.[74] According to the objective criterion, data such as IP addresses may be regarded as being personal data in the hands of an entity such as the Federal Republic of Germany 'even if only' a third party (the ISP) is able to determine the identity of the data subject. According to the relative criterion:

> such data may be regarded as personal data in relation to an entity such as Mr Breyer's internet service provider because they allow the user to be precisely identified . . ., but not being regarded as such with respect to another entity, since that operator does not have, if Mr Breyer has not disclosed his identity during the consultation of those websites, the information necessary to identify him without disproportionate effort.[75]

In *Breyer*, the CJEU did not opt for the objective criterion but seems to have applied a modified or more nuanced version of the relative criterion. The reference to the means likely reasonably to be used by both the controller and by 'any other person' in Recital 26 DPD, the Court held, suggests that, for information to be treated as personal data, 'it is not required that all the information enabling the identification of the data subject must be in the hands of one person'.[76] However, this did not 'automatically' make the data in the hands of a party which, like the Federal Republic of Germany in *Breyer*, did not have the means in its hands to identify its users, personal data. As Advocate General Campos Sánchez-Bordona explained:

> That overly strict interpretation would lead, in practice, to the classification as personal data of all kinds of

70 Ibid 9.

71 See Ohm (n 48) 1704. El Emam and Ávarez are also critical of the zero-risk approach of WP 216 and claim that 'this will not work in practice'. See Khaled El Emam and Cecilia Ávarez, 'A Critical Appraisal of the Article 29 Working Party Opinion 05/2014 on Data Anonymization Techniques' (2015) 5 International Data Privacy Law 73.

72 Ibid 1753.

73 Case C-582/14 *Patrick Breyer v Bundesrepublik Deutschland* [2016] ECLI:EU:C:2016:779, para 27.

74 See Opinion of AG Campos Sánchez-Bordona in Case C-582/14 *Patrick Breyer v Bundesrepublik Deutschland* [2016] ECLI:EU:C:2016:339, para 52 and para 53, and CJEU in *Breyer* (n 73) para 25.

75 *Breyer* (n 73) para 25.

76 Ibid para 43.

information, no matter how insufficient it is in itself to facilitate the identification of a user.[77]

The analysis, as the CJEU explained, has to be more nuanced and one has to investigate 'whether the possibility to combine a dynamic IP address with the additional data held by the internet service provider constitutes a means likely reasonably to be used to identify the data subject'.[78] That would *not* be the case, as the CJEU further explained, basing itself on the Advocate General's opinion,[79] in the following two situations: (i) in cases where the identification of the data subject was prohibited by law, or (ii) if identification was 'practically impossible on account of the fact that it requires a disproportionate effort in terms of time, cost and manpower, so that the risk of identification appears in reality to be insignificant'.[80] Though the CJEU seems to set the bar very high when it requires the risk of re-identification to be 'insignificant', the court's reference to 'practical impossibility' rather than to 'impossibility' appears to us to be a clear affirmation of the risk-based approach and a negation of the strict approach of WP 216. It also, in fact, reiterates some of the objective factors specified in Recital 26 GDPR. The CJEU therefore concluded in *Breyer* that the data in the German Federal Republic institutions' hands were personal data, but only because of the existence of legal channels which enabled the Federal Government to ask the competent authority to obtain identifying information from the ISP in the event of a cyber-attack.[81] In the absence of these channels, the data would not have been considered personal data simply because a known third party could identify them.

The judgment of the CJEU in *Breyer* is important authoritative support for the view that Recital 26 DPD and, by extension, Recital 26 GDPR, points towards a risk-based approach to the notion of personal data and not the strict approach of WP 216. The fact that the phrase 'or by any other person' in Recital 26 DPD has been changed to 'or by another person' in Recital 26 GDPR does not alter our view. Dalla Corte has claimed that that this change in wording may be more than cosmetic and signals an intention to narrow the ambit of the personal data concept by restricting the range of 'third parties who may be approached by a controller to identify the data subject . . . [to] the ones that can reasonably be accosted.'[82] We hold, however, that although the change in wording may very well lead to a restriction of legally relevant agents of identification that 'controllers' may have recourse to, it does not alter the fact that the preposition 'or' maintains the range of potential actors wide since 'another person' could be a third party,[83] such as an intruder who, whether intentionally or inadvertently and without recourse to illegal means, may have successfully been able to re-identify the data subject.[84]

Although, as evidenced by Article 29(1) DPD, the Working Party's role under the DPD was advisory, its opinions have been highly influential in data protection practice as they shed light on how data protection authorities interpret and are likely to enforce data protection law. Nevertheless, it appears that some individual data protection authorities have not taken the rigid approach of WP 216. The UK Information Commissioner's Office adopted a more balanced understanding of Recital 26 DPD in its anonymization code of practice of 2012,[85] a view that it appears to maintain in its draft anonymization guidance in its version of May 2021.[86] The Irish Data Protection Authority 'does

77 See Opinion of AG Campos Sánchez-Bordona in *Breyer* (n 74) para 65.

78 *Breyer* (n 73) para 45.

79 See Opinion of AG Campos Sánchez-Bordona in *Breyer* (n 74) para 68.

80 *Breyer* (n 73) para 46.

81 Ibid, para 47.

82 L Dalla Corte, 'Scoping Personal Data: Towards a Nuanced Interpretation of the Material Scope of EU Data Protection Law' (2019) 10(1) European Journal of Law and Technology 1, 15.

83 A 'third party' is, after all, any person other than, inter alia, the controller, cf art 4(10) GDPR.

84 Bygrave and Tosoni also doubt whether the difference in wording leads to a significant change of 'personal data' since the difference concerns a recital and not the definition of 'personal data', making the CJEU unlikely to depart substantially from its DPD-era rulings on the definition. As these authors note, their scepticism is supported by Advocate General Bobek in FashionID as: 'as Article 4 of the GDPR largely retains the same key terms as Article 2 of Directive 95/46 . . ., it would be rather surprising if the interpretation of such key notions, including the notion of . . . personal data, were to significantly depart (without a very good reason) from the extant case-law'. See Case C-40/17, Fashion ID (AG Opinion), para 87 and Bygrave and Tosoni (n 25) in C Kuner, LA Bygrave and C Docksey, (eds), *Update of Selected Articles to The EU General Data Protection Regulation: A Commentary* (OUP Oxford 2021) 25.

85 Annex 1 to the anonymization code contains a case study whereby research data held by the University of Stevenham Research Centre (USRC) is redacted by USRC and disclosed to a neighbouring research centre (NRC) following a freedom of information request from NRC. According to the code, 'The redacted data-set is still personal data in the hands of USRC because it still holds the full version of the original research data. This could act as a "key" that would allow the extracted data to be linked back to personal identifiers . . .. The extract is only be [sic] personal data in the hands of USRC because only USRC holds the "key" needed to make the link back to the personal identifiers it holds. NRC cannot do this because there is no information in the extract itself that could allow the linkage to be made. This shows that at the point at which USRC discloses the extract, it ceases to be personal data – even though it is still personal data in the hands of USRC as long as it holds "the other information" necessary to enable identification.' See UK Information Commissioner's Office, *Anonymisation: Managing Data Protection Risk - Code of Practice* (UK ICO 2012) 58–59.

86 At the time of writing, the ICO is updating its 2012 Code of Practice and has announced that, as from May 2021, it will be publishing draft chapters of its Anonymization, pseudonymization, and privacy enhancing technologies guidance. According to the draft guidance: 'In the ICO's

not deem it necessary to prove that it is impossible for any data subject to be identified in order for an anonymisation technique to be considered successful'.[87] On similar lines and reminiscent of the CJEU in Breyer, the French Commission Nationale de l'Informatique et des Libertés (CNIL) describes anonymisation as processing which consists in an ensemble of techniques which render identification of the data subject 'practically impossible'.[88]

We return to this issue in the section 'Discussion and way forward' of this article.

Case studies

We now review two case studies related to the anonymization of unstructured data. The case studies respectively focus on text documents and medical images. Both studies provide empirical support to the central claim of this article, namely that anonymization of unstructured data that satisfies the criteria of WP 216 is impossible to achieve in practice, unless the data controller decides to delete the original dataset.

Case study 1: Anonymization of court cases

In this first case study, we investigate the extent to which GDPR-compliant anonymization of textual documents is technically possible. The dataset that will be used for this case study is a collection of 13,759 court cases from the European Court of Human Rights (ECHR), which are made available on the web portal of the Court. Those court cases include detailed information in plain text about various individuals (name, date of birth, criminal record, family status, etc.) mentioned in the court cases. This information does not only relate to the plaintiffs, but also to various parties involved in the case, such as witnesses, lawyers, judges, and government agents. The choice of ECHR court cases for this case study is motivated by practical considerations, as ECHR court cases provide a convenient and publicly available resource for investigating how personal information is expressed in text documents.[89]

The bulk of the personal information provided in ECHR court cases can be found in the section 'Circumstances of the case' that introduces the factual elements that underlie the application. We provide an example of such section in Table 2 (top part). Notwithstanding the name of the applicant himself, the text of the case contains various types of information (such as place and date of birth, dates of various judgments) that, taken together and combined with external knowledge sources, makes it possible to re-identify the individual in question. For instance, the combination of the person's place and year of birth (1955 in Bridgend) with the date of death of his spouse (29 April 1999) will narrow down the set of possible individuals to a single person.

Manual de-identification

A first attempt to address this re-identification risk is to remove from the text all direct and indirect identifiers that may be related to the individual. To investigate how this masking process can be conducted in practice, we hired a group of 12 law students from the University of Oslo and asked them to read through a collection of ECHR court case and subsequently mark within each case all text spans that may directly or indirectly contribute to the re-identification risk. The text spans masked by the students were grouped in 8 distinct categories:

- Names of individuals

view, the same information can be personal data to one organisation, but anonymous information in the hands of another organisation. Its status depends greatly on its circumstances, both from your perspective and in the context of its disclosure.' See UK Information Commissioner's Office, *Introduction to Anonymisation: Draft Anonymisation, Pseudonymisation and Privacy Enhancing Technologies Guidance* (May 2021) 9. At the time of writing, the draft is open for consultation. See <http://ico.org.uk/about-the-ico/ico-and-stakeholder-consultations/ico-call-for-views-anonymisation-pseudonymisation-and-privacy-enhancing-technologies-guidance/> accessed 4 June 2021. Following Brexit, after 31 January 2020, the GDPR no longer applies in the UK. Although the UK has enacted its own version of the GDPR whose articles, at the time of writing, are largely similar to those in the EU's GDPR, (eg the definition of personal data is identical), the UK ICO's pronouncements are important because they show that even after the publication of WP 216 in 2014, the ICO maintained its relativist understanding of the identifiability test.

Such a perspective has also been favoured by the UKAN since 'it directly ties the concept of anonymization to the notion of the context of the personal data'. See Elliot and others (n 47) 24. See also <http://www.ukanon.net> accessed 4 June 2021.

87 Irish Data Protection Commission, Guidance Note: Guidance on Anonymization and Pseudonymization (2019) 5 <www.dataprotection.ie/sites/default/files/uploads/2019-06/190614%20Anonymisation%20and%20Pseudonymisation.pdf> accessed 4 June 2021.

88 See CNIL, 'L'anonymisation de données personelles' (2020) <www.cnil.fr/fr/lanonymisation-de-donnees-personnelles> accessed 4 June 2021. ('Anonymisation est un traitement qui consiste à utiliser un ensemble de techniques de manière à rendre impossible, en pratique, toute identification de la personne par quelque moyen que ce soit et de manière irréversible.') Emphasis added. See also *Breyer* (n 73) para 46.

89 Of course, conducting a complete anonymization of ECHR court cases is probably not a very useful operation to perform in practice (as court cases are generally meant to be published or at least made available in some form to legal professionals), but this case study is only meant for illustrative purposes. The extent to which the GDPR applies to the ECHR's database is also unclear due to the ECHR's international law immunities, see Christopher Kuner, 'International Organizations and the EU General Data Protection Regulation', University of Cambridge Faculty of Law Research Paper No 20/2018, February 2018 <https://ssrn.com/abstract=3050675> accessed 10 September 2021.

Table 2. Excerpt from an ECHR court case [nr 61391/00]: original (top), de-identified version after masking direct and indirect identifiers (middle), fully anonymised version where all phrases that can potentially link back to the original document are masked (bottom)

1. The applicant [Mr Colin Joseph O'Brien] was born in 1955 and lives in Bridgend.
2. His wife died on 29 April 1999 leaving two children, born in 1989 and 1991.
3. In 1999 the applicant enquired about widows' benefits and he was informed that he was not entitled to such benefits.
4. In early 2000 the applicant applied for widows' benefits again and on 13 March 2000 the Benefits Agency rejected his claim.
5. He lodged an appeal against this decision on 16 March 2000 and this appeal was struck out on 23 May 2000 on the basis that it was misconceived.
6. On 16 May 2000 the applicant made an oral claim for Widow's Bereavement Allowance to the Inland Revenue. On 23 May 2000 he was informed that his claim could not be accepted because there was no basis in domestic law allowing widowers to claim this benefit. The applicant was advised that an appeal against this decision would be bound to fail.
7. The applicant received child benefit in the sum of GBP 100 per month.

1. The applicant [***] was born in *** and lives in ***
2. His wife died on *** leaving *** children, born in ***
3. In *** the applicant enquired about widows' benefits and he was informed that he was not entitled to such benefits.
4. In *** the applicant applied for widows' benefits again and on *** the *** rejected his claim.
5. He lodged an appeal against this decision on *** and this appeal was struck out on *** on the basis that it was misconceived.
6. On *** the applicant made an oral claim for Widow's Bereavement Allowance to the Inland Revenue. On *** he was informed that his claim could not be accepted because there was no basis in domestic law allowing widowers to claim this benefit. The applicant was advised that an appeal against this decision would be bound to fail.
7. The applicant received child benefit in the sum of *** per month.

1. The applicant [***] was born in *** and lives *** ***
2. *** *** *** *** *** *** two *** *** *** *** *** *** ***
3. In *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** was *** *** *** *** *** ***
4. In *** the applicant *** *** *** *** *** *** *** *** *** the *** *** his *** ***
5. *** *** an *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** the *** that it was *** ***
6. *** *** *** *** *** *** *** *** for *** *** *** *** the *** *** *** *** *** *** *** *** *** *** *** *** could *** *** *** *** *** *** *** *** *** in *** law *** *** to *** this *** *** *** *** *** *** *** *** *** *** *** this *** *** *** *** to *** ***
7. The *** *** *** *** in the *** *** *** *** *** ***

- Names of organizations (companies, public institutions, etc.)
- Places and geographical locations
- Date and time indicators
- Demographic attributes (ethnicity, age, gender, etc.)
- Quantities (monetary values, number of convictions, etc.)
- Codes (application number, phone number, etc)
- Miscellaneous direct or indirect identifying information (not belonging to any of the above categories)

The result of this manual de-identification process is illustrated in Table 2 (middle part). As we can observe, the edited version of the court case is stripped of all direct and indirect identifiers that may single out the identity of the applicant. The only information that we can gather about the applicant from the text is that the person is male, most likely a British national (due to references to British public agencies such as the Inland Revenue), a widower with children, and has been denied widows' benefits at an undisclosed point in time. We contend that, in the absence of other information related to the application, those pieces of information are on their own insufficient to single out the identity of that person.[90]

90 It should be stressed again that the use of ECHR court cases in this case study is only meant for illustrative purposes in order to highlight the concrete challenges involved in anonymizing text data. In practice, such court cases are typically made available in a broad range of publication

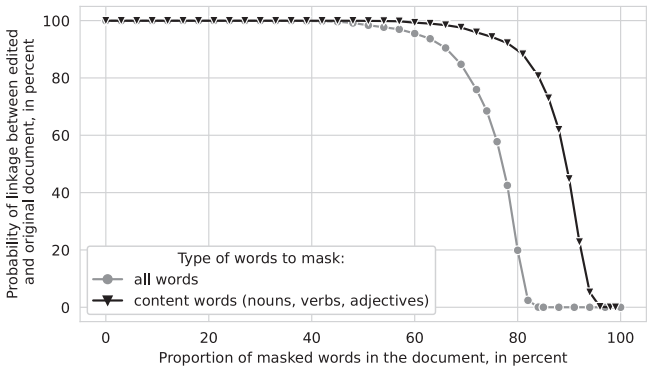

Figure 1: Relation between the proportion of masked words and the probability of linkage between the edited and original document, as computed from the dataset of 13,759 court cases from the ECHR. The line denoted by triangles indicates how the probability of linkage evolves in relation to the proportion of masked words (taking all words into account) in the edited document, while the line denoted by circles focuses specifically on the proportion of content words (defined here as nouns, verbs and adjectives).

Linkability with original dataset

Is this process of stripping direct and indirect identifiers from the text sufficient to ensure anonymization that complies with the criteria set forth in WP 216? No, as the edited documents can still be linked back to the original documents through a relatively simple process. If we assume that the data controller who carried out the de-identification procedure retains the original (unedited) collection of documents (which they typically do), such data controller can relatively easily search for the remaining 'text phrases' that occur in the edited document to determine its most likely original source.

For instance, in the collection of 13,759 court cases employed in this case study, only one single document contains the phrase 'was advised that an appeal against' which occurs in the above example. Even short phrases can link back to the original document when one searches for their occurrence within the same document. For instance, although the phrases 'rejected his claim' and 'could not be accepted' can both be found in several documents, their occurrence within the same document can again only be observed in one single court case.

Can we filter out those phrases to ensure that no edited document can be traced back to its original version? Yes, but this requires the removal of most of the document's content, which renders the anonymization process essentially meaningless.

To corroborate this hypothesis, we developed a simple software system[91] that takes a collection of documents as input and searches for all phrases that, isolated or in combination, can link the edited document back to its original.

The result of such a thorough anonymization process is illustrated in Table 2 (bottom part). As we can observe, the resulting text has essentially been stripped of all useful content, and only a few generic words such as prepositions, determiners and common phrases such as 'the applicant' are preserved.

This large proportion of masked tokens is not specific to the example in Table 2. After running this automated anonymization process on all 13,759 court cases employed in this case study, we find that the requirement of non-linkability between the edited and original document leads to the masking, on average, of at least 80 per cent of the words occurring in each court case. Furthermore, if we concentrate on so-called 'content words', which are words such as nouns, verbs and adjectives (but not function words such as prepositions or articles), this proportion of masked words increases to 96 per cent, as illustrated in Figure 1. To put it differently, the requirement of non-linkability not only leads to removing most words from a court case, but it also removes the most important ones in terms of data utility. This phenomenon can be clearly observed in the masked text from the bottom part of Table 2, where the only remaining content words are 'applicant', 'born', 'lives' and 'law', out of a total of 172 words in the original excerpt.

Although the results presented above are estimated from a specific collection of court case documents, they are not surprising from the perspective of corpus

channels and are heavily cross-referenced, which would make their anonymization ineffective for most purposes.

91 The source code for this software is available by request. Technically speaking, the implementation of this software rests on the construction of a special data structure called an 'inverted index'. An inverted index maps each possible word or phrase to the documents that include it. Using such an index, one can easily infer the list of phrases that, individually or in combination with other phrase, lead to a unique document. Once extracted, the resulting phrases can be masked from their corresponding documents.

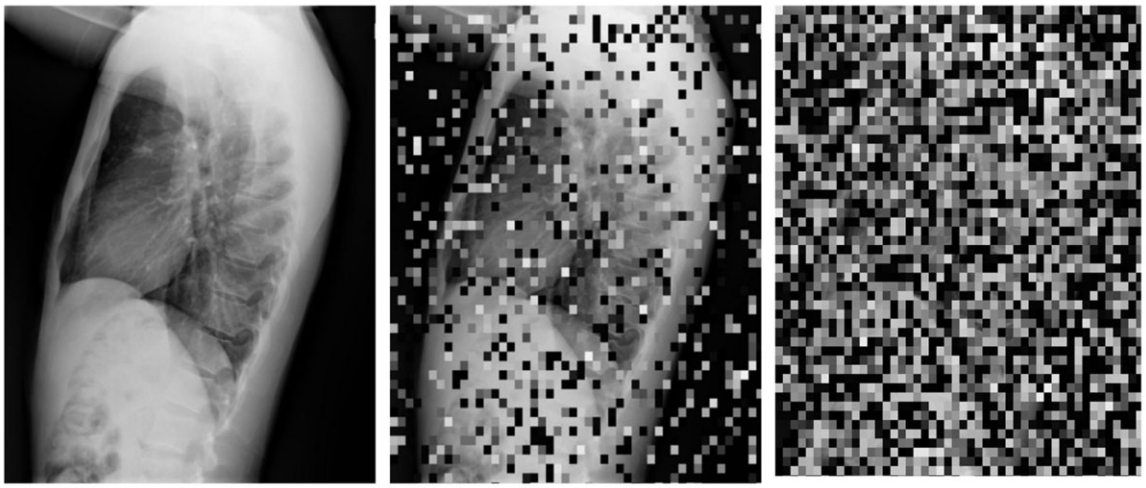

Figure 2: Original image of chest X-ray sampled from the dataset (left) and corresponding images after blurring respectively 30% (middle) or 85% (right) of the X-ray.

linguistics. Indeed, the proportion of phrases occurring only once in a collection of texts (a phenomenon also called *hapax legomenon*) increases rapidly with the size of those phrases. This also holds for very large document collections and for web-scale data.[92] In other words, most phrases of more than 5–6 words will occur only once in a collection of documents and can therefore be exploited to reconnect de-identified documents with their original source.

This case study illustrates that, for textual data, WP216-compliant anonymization needs to go far beyond the mere removal of direct and indirect identifiers. Indeed, the requirement of non-linkability between the original and edited document makes it essentially impossible to anonymise text data without rendering the resulting text essentially useless for most practical purposes, at least if we assume that the data controller retains a copy of the original dataset (which is typically the case).

Case study 2: Anonymization of medical images

The findings obtained in the above case study are not specific to text data and can also be reproduced for other types of unstructured data such as images, audio recordings and videos. In this second case study, we concentrate on medical images. In addition to being personal data, medical images are also a specific type of 'data concerning health', cf Article 4(15) GDPR, and are therefore special categories of personal data pursuant to Article 9(1) GDPR.

Our dataset consists of a total of 7570 chest X-ray images published by the US National Library of Medicine and made publicly available for research purposes.[93] The images do not contain any identifier associated with the patient, making it impossible to determine the identity of the patients without additional information. We illustrate one example of such X-ray image in Figure 2 (left).

Although none of the patients can be identified from those chest X-rays alone, the dataset does not qualify as being anonymous according to the GDPR requirements. Indeed, the healthcare institution from which the image originates has most likely kept an archive of those X-ray images together with meta-data including the patient identity. Consequently, the data controller (or any individual that has access to the original database) can conduct an 'image search' to link the images contained in the datasets with such meta-data, and therefore determine the identity of the patient.

It is, however, possible to introduce artificial noise in the image to increase the difficulty of executing such a linkage. As with case study 1, we analysed the level of noise necessary to reduce the probability of linkage to a number close to zero. The artificial noise introduced in

92 See Joaquim F Silva and Jose C Cunha, 'An Empirical Model for n-gram Frequency Distribution in Large Corpora', Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, 11–14 May 2020, Proceedings, Part II, 12085, 840–851.

93 D Demner-Fushman and others, 'Preparing a Collection of Radiology Examinations for Distribution and Retrieval' (2016) 23(2) Journal of the American Medical Informatics Association 304–310.

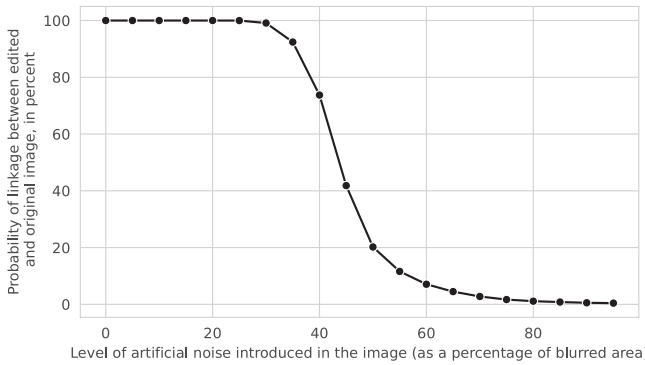

Figure 3: Relation between the level of artificial noise introduced in the image and the probability of linkage between the edited image and the original X-ray.

this experiment consisted in blurring local regions of the image. The middle and right-hand side of Figure 2 shows edited versions of the image respectively blurring 30 per cent or 85 per cent of the original X-ray.

The relation between the level of introduced noise and the probability of linkage through image search is illustrated in Figure 3. As can be observed from the figure, the probability of linkage can only be reduced to a low level on the condition of introducing levels of noise above 85 per cent. This corresponds to blurring the image as in the right-hand side of Figure 2.

Such levels of artificial noise render the image essentially useless for most practical purposes, and certainly for applications such as medical research (as even basic anatomical features become unrecognizable). In other words, when interpreted in a strict sense, the requirement of non-linkability between the original data from the data controller and the released dataset means that the WP 216-compliant anonymization of images and other visual media becomes essentially impossible.

Discussion and way forward

Resolving the conflicting interpretations

The section 'Case studies' shows the dramatic consequences if one applies the strict approach in WP 216 to unstructured data. For the resultant data to be categorized as anonymized data, the original data must be deleted after the de-identification procedure. This is irrespective of whether the anonymized data is to be released to a third party or to be kept by the same data controller, for example, by a different department than that holding the original data. Where the original data controller wants or is legally bound to keep a copy of the original data set, the anonymized data that is produced—if it is to meet the stringent requirements of WP 216—is practically useless.[94] Hence our claim in the introduction to this paper that it is virtually impossible to anonymize unstructured data in a manner that satisfies the requirements of WP 216.

This empirical finding sharpens the legal discussion of whether the risk-based or strict approach constitutes the most defensible interpretation of anonymization requirements. In our view, the approach to the resolution of these conflicting interpretations can and should be resolved through the standard interpretive techniques in EU law, namely textualism, context, and teleology.[95] It cannot be resolved by simply referring to WP 216. While the European Court of Justice has cited at times the opinions of the working party as persuasive authority, they have no legal weight and remain only guidelines and are not legally binding.[96] Moreover, the EU Commission has underlined the soft law nature of the WP 216's work in its 2018 communication to the

94 See Ohm (n 48) 1704. As Elliot and others state: 'Zero risk is not a realistic possibility if you are to produce useful data.' Elliot and others *The Anonymisation Decision-Making Framework* (n 47) 19, 33 and 34. See also Elliot and others, 'Functional anonymisation: Personal data and the data environment' (n 53).

95 Koen Lenaerts and Jose Gutierrez-Fons, 'To Say What the Law of the EU is: Methods of Interpretation and the European Court of Justice' (2014) 20(2) The Columbia journal of European Law 3.

96 Adam Finlay, 'Is the Art29 Working Party Undermining Itself?' (*Lexology*, 30 January 2018) <https://www.lexology.com/library/detail.aspx?g=aa713293-f9ef-496f-b4e3-5f5a8b5ece10> accessed 10 March 2022. Pursuant to art 70(1)(e) GDPR, the EDPB nowadays continues the advisory role of its predecessor by issuing guidelines, recommendations, and best practices which, though having a great deal of influence, are not directly binding on third parties. See Christopher Kuner, 'Territorial Scope and Data Transfer Rules in the GDPR: Realising the EU's Ambition of Borderless Data Protection' (16 April 2021) University of Cambridge Faculty of Law Research Paper No 20/2021, 32, <https://ssrn.com/abstract=3827850> or <http://dx.doi.org/10.2139/ssrn.3827850> accessed 10 March 2022.

Parliament and Council: 'where questions regarding the interpretation and application of the Regulation arise, it will be for courts at a national and EU level to provide the final interpretation of the Regulation'.[97]

When it comes to *textualist* interpretation, it is arguably difficult to distinguish between the claims of the risk-based and strict approaches. Both are plausible interpretations as to what constitutes a 'reasonable risk' of identification of personal data in the context of anonymization. In the event though that a literal interpretation does not resolve the meaning of a provision, it is incumbent to look to context and purpose.[98] As to *context*, the presumption in the context of the EU is that that the 'legislator is a rational actor. This means that the authors of the Treaties are assumed to have established a legal order that is consistent and complete'.[99] In our view, the strict approach in WP 216 is difficult to reconcile with Article 2(2) of the Free Flow of Non-Personal Data Regulation discussed earlier in this article.[100] The premise underlying that article is that it is sometimes possible that the personal data and the non-personal data in a mixed dataset may be separated, in which case the Free Flow of Non-Personal Data Regulation applies to the non-personal data part of the data set, whereas the GDPR applies to the personal data part of the data set. Unstructured text documents are perhaps the quintessential example of mixed datasets, since they typically contain both personal data and non-personal data. Where it is technically possible to remove personal data from a mixed dataset, can the original mixed dataset continue to co-exist alongside the separate sets of personal and non-personal datasets? It would seem that once non-personal data have been extracted from a mixed dataset, in order for that extracted non-personal data to continue to exist as non-personal data, the original mixed dataset would have to be rendered virtually useless. This would considerably restrict the utility—the *effet utile*—of the Free Flow of Non-Personal Data Regulation since few data controllers are likely to be interested in deleting the original mixed dataset.[101]

In relation to *teleological* approaches, they are central in the interpretation of much EU law given the abstract nature of drafting but also its 'purpose-driven functionalism' in order to secure 'objectives of paramount constitutional importance'.[102] Teleological interpretation has at least three different modes:[103](i) securing an *effective* interpretation in light of the legal context; (ii) ensuring when there is ambiguity that the interpretation is in line with the *objectives* the provision seeks to pursue; (iii) and avoiding absurd *consequences* of literal interpretation. When there is a clash between objectives, a proportionality approach is to be adopted as to which one should prevail.

Let us take each in turn. First, the risk-based approach is arguably a more effective approach in reflecting the patchwork of legal provisions that form the context for interpretation. This applies not only to ensuring consistency with the Free Flow of Non-Personal Data Regulation, but also to at least Recital 26 GDPR and Article 25(1). The latter require inter alia periodic assessment of the state-of-the art technology (including technological measures) that may be used. If, as according to these legal sources, interpretation of reasonable risk can change according to the available technology, an interpretation that renders all anonymization techniques to be outside the four walls of the GDPR makes these clauses effectively redundant.

Second, the purpose of the GDPR is not to eliminate all risks. As evidenced from the second and third paragraphs of Article 1, the objective of the GDPR is both to protect fundamental rights, in particular, the right to data protection, as well as to ensure that the free movement of personal data within the EU 'shall be neither restricted nor prohibited'. As Article 29 Working Party itself acknowledged, the risk-based approach gained much more attention in the discussions at the European

97 Stronger protection, new opportunities—Commission guidance on the direct application of the General Data Protection Regulation as of 25 May 2018, COM(2018) 43 final.

98 Case C-220/03, *European Cent. Bank v Fed. Republic of Germany*, 2005 E.CR 1-10595, para 31. ('Article 8(1) of the Agreement expressly and unambiguously makes the refund of turnover tax subject to the condition, not fulfilled in the present case, that that tax be "invoiced separately". Although an interpretation of a provision of an Agreement "in the light" of its legal context is possible in principle to resolve a drafting ambiguity, such an interpretation cannot have the result of depriving the clear and precise wording of that provision of all effectiveness.'). See also Case C-48/07, *Belgium v Les Vergers du Vieux Tauves*, 2008 ECR 1-10627, para 44 ; Case C-263/06, *Carboni e Derivati Srl v Ministero dell'Economia e delle Finanze and Riunione Adriatica di Sicurta SpA*, 2008 ECR 1-1077, para 48.

99 Lenaerts and Gutierrez-Fons (n 95) 17. See, for example, the importance of reading together internal provisions in Case C-465/07, *Meki Elgafaji and Noor Elgafaji v Staatssecretaris van Justitie*, 2009 ECR 1-921, 28-2.9.

100 See the section 'Anonymous data as the antithesis of 'Personal Data'' of this article. See also Elliot and others (n 47) 33.

101 On *effet utile*, see Urska Sadl, 'The Role of Effet Utile in Preserving the Continuity and Authority of European Union Law: Evidence from the Citation Web of the Pre-accession Case Law of the Court of Justice of the EU' (2015) 8(1) European Journal of Legal Studies 18–45.

102 Lenaerts and Gutiérrez-Fons (n 95) 31.

103 Joxerramon Bengoetxea, *The Legal Reasoning of the European Court of Justice: Towards a European Jurisprudence* (Oxford, Clarendon Press 1993).

Parliament and at the Council during the legislative process of the GDPR, was introduced as a core element of the accountability principle itself (Article 24, cf Article 5(2)), and has been extended and reflected in 'inter alia' the obligation of security (Article 32), the obligation to carry out an impact assessment (Article 35), the data protection by design principle (Article 25) and the obligation for documentation (Article 30).[104]

One can also question whether the strict interpretation of WP 216, which requires anonymized data to be impossible to link back to their original source, leads to a better privacy protection for the data subjects. We contend that this is not the case when one wishes to share anonymized data with a third party (that is, when the original dataset is held by one controller and the anonymized data is held by a different controller).[105] Indeed, this WP 216 requirement focuses on preventing linkage with a dataset (the original, unedited data) that is not available to a motivated intruder with access to the anonymized data, and is therefore a risk of little relevance when it comes to protecting the privacy of the individuals whose data has been anonymized in the dataset.[106] This is notably illustrated in the first case study described in this article, where the requirement of non-linkability leads to the removal of numerous phrases such as 'rejected his claim' and 'could not be accepted', although those phrases did not offer any (direct or indirect) information about the identity of the individual in question.

Third, a risk-based approach avoids the draconian and absurd consequences of the strict approach. By providing an evaluation in the context of an actual data environment, it offers a more nuanced and practical alternative to the rather rigid approach of WP 216, with its almost absurd implications, as demonstrated in the two case studies described in this article. As Elliot and others explain, it is a misapprehension 'that anonymisation can be absolute without mangling the data so badly that it has no utility whatever':[107]

> If anonymisation is to be a useful tool for data and risk management, one has to specify its circumstances. Thus the only sensible response to the question 'are these personal data?' is another question: 'in what context?' or more specifically 'in what data environment?'[108]

Drawing together this analysis, it is arguable that by introducing legal context and a teleological perspective the risk-based approach becomes the most plausible and defensible interpretation. Anonymization of personal data is possible but the relevant risks must be considered for each case. To be sure, introducing these contextual and teleological approaches to anonymization leads to the result that the GDPR could be under-applied rather than over-applied—as more data would not be classified as personal data. This is somewhat paradoxical and contrary to the common result of using these interpretive methods, which is often more expansive and 'activist' applications of EU law. However, such a restrictive outcome is not uncommon. As Lenaerts and Gutierrez-Fons, point out, in *Kalfelis*,[109] the 'ECJ engaged in a teleological reduction of the scope of Article 6(1) of the 1968 Brussels Convention'.[110]

The issue of temporality

Changes in technology are, however, a two-edged sword for the risk-based approach. On one hand, they underscore the theoretical argument for the risk-based approach—as is argued in the section 'Resolving the conflicting interpretations'. As discussed earlier,[111] temporality is one of the objective factors mentioned in Recital 26 GDPR. On the other hand, improvements in identification technology limit the practical application of the risk-based approach. Re-identification risks do not necessarily stay constant over time, and may evolve due to technological advances or as a result of the increasing availability of online data that may be collected on various individuals. In addition, they solidify the absolutism of the strict approach. As described in this article, WP 216 requires the complete and irreversible removal of all identifiers (direct and indirect) from a dataset for it to be deemed anonymized. This irreversibility should not only consider re-identification risks that may occur in the present state of technology but also those that may occur in the future.

104 Article 29 Working Party, 'Statement on the role of a risk-based approach in data protection legal frameworks' (WP 218, 30 May 2014), 2.

105 For the data controller of the original, unedited dataset (or for a motivated intruder that has obtained access to the data controller's infrastructure), performing such a linkage is technically possible, but would be a meaningless operation, as it would simply amount to recreating the exact same data they already have.

106 As Esayas states, '[i]t is true that re-identification has become easy as a result of the technological advancement and the ubiquity of information on the Internet, but the alternative should not be a boundless and overboard application of the [Data Protection] Directive.' See Samson Yoseph Esayas, 'The Role of Anonymisation and Pseudonymisation under the EU Data Privacy Rules: Beyond the "all or nothing" Approach' (2015) 6 (2) European Journal of Law and Technology 7.

107 Elliot and others (n 47) 52.

108 Ibid.

109 Case 189/87, *Kalfelis v Bankhaus Schroder and Others*, 1988 ECR 5565.

110 Lenaerts and Gutierrez-Fons (n 95) 37.

111 See the section 'The risk-based approach' of this article.

This question of temporality is an important and difficult one. One of the most important risks that data controllers need to consider when processing unstructured data are the expected technological advances in Artificial Intelligence (AI), especially within the fields of natural language processing, computer vision and speech processing. For instance, AI-based models relying on a wide range of stylistic patterns may be applied to predict with relatively high precision the author of short texts.[112] Similarly, it has been shown that images containing human faces can in certain conditions be automatically 'deblurred' using advanced computer vision models.[113] The identity of speakers in audio recordings can sometimes also be retrieved even after the application of obfuscation methods to disguise or distort their voices.[114]

This technological development, however, proceeds in both directions. The topic of privacy-preserving techniques is one of the most active research areas within Artificial Intelligence,[115] and recent years have seen the development of various adversarial AI-models that are specifically optimized to be robust to privacy attacks and prevent the leak of personal information either from data releases[116] or from machine learning models trained from those.[117] This development also applies to unstructured data such as texts or images, although several open research questions remain.[118]

Compared to structured datasets, unstructured data are typically more difficult to link to other data sources (other than the original dataset), as they lack a predefined tabular structure making it possible to directly merge several data sources based on shared variables. However, one can rely on information extraction techniques to automatically derive structured representations from text documents,[119] and subsequently use those structured representations as a starting point to link the documents to other data sources. Such automated extraction of structured representations can also be performed on other types of unstructured data such as images.[120]

While data may be considered safely anonymized at the time of its release, there is therefore always a residual risk that the data may be re-identified at some point in the future, due to technological advances and the growth of online data that can be collected on individuals.[121] The data controller should thus monitor the data environment once data has been shared or otherwise disclosed as this is not least invaluable later when the controller is considering the next release.[122]

Way forward and role of the EDPB in providing clarity

In a February 2021 advisory document, the EDPB—Article 29 Working Party's successor—stated that '[t]he determination of whether information is anonymous must be made by the application of the test of identifiability outlined in Recital 26 GDPR'.[123] It emphasized that all of the factors in this recital 'must be considered in making an assessment as to the reasonable likelihood of identifiability'.[124] However, the EDPB also stated that WP 216 'should be taken into account',[125] and that:

> [a]ny such assessment should be made along the lines suggested by the CJEU in *Breyer*, which refers to Recital 26 of Directive 95/46/EC, looking at the legal and practical means by which re-identification may be effected by the use of additional data in the hands of third parties.[126]

Rather than add clarity, this makes the situation arguably more nebulous.[127] It is perhaps trite to note that

112 See P Shrestha and others, 'Convolutional Neural Networks for Authorship Attribution of Short Texts' in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics* (2017) 669–674.

113 Z Shen and others, 'Exploiting Semantics for Face Image Deblurring' (2020) 128(7) International Journal of Computer Vision 1829–46.

114 M Farrús, 'Voice Disguise in Automatic Speaker Recognition' (2018) 51(4) ACM Computing Surveys (CSUR) 1–22.

115 See eg M Gong and others, 'A Survey on Differentially Private Machine Learning' (2020) 15(2) IEEE Computational Intelligence Magazine 49–64.

116 S Shaham and others, 'Privacy Preserving Location Data Publishing: A Machine Learning Approach' (2020) *IEEE Transactions on Knowledge and Data Engineering*.

117 T Xiao and others, 'Adversarial Learning of Privacy-preserving and Task-oriented Representations' (2020) 34(7) Proceedings of the AAAI Conference on Artificial Intelligence 12434–41.

118 See P Lison and others, 'Anonymisation Models for Text Data: State of the Art, Challenges and Future Directions' *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)* (2021).

119 See eg JL Martinez-Rodriguez, A Hogan and I Lopez-Arevalo, 'Information Extraction Meets the Semantic Web: A Survey' (2020) Semantic Web 1–81.

120 B Zhou and others, 'Semantic Understanding of Scenes through the Ade20k Dataset' (2019) 127(3) International Journal of Computer Vision 302–21.

121 See Elliot and others (n 47) 128.

122 Elliot and others provide examples of various measures that the data controller may take to monitor the data environment once the data has been shared. Elliot and others (n 47) 128.

123 See EDPB, 'EDPB Document on response to the request from the European Commission for clarifications on the consistent application of the GDPR, focusing on health research' (n 61) para 45.

124 Ibid para 46.

125 Ibid.

126 Ibid.

127 In fact, the Multistakeholder Expert Group to the Commission 2020 evaluation of the GDPR states that several members of the group seek more guidance from the EDPB on inter alia anonymization, following a risk-based methodology. See Multistakeholder Expert Group, 'Report – Contribution from the Multistakeholder Expert Group to the Commission 2020 Evaluation of the GDPR' (17 June 2020 <https://ec.europa.eu/info/sites/info/files/report_from_multistakeholder_expert_group_on_gdpr_application.pdf>) accessed 10 March 2022.

'you cannot have your cake and eat it'. One cannot apply 'both' the strict approach of WP 216 and, at the same time, the risk-based approach of Recital 26 GDPR. Though the CJEU in *Breyer* seems to set a high bar in requiring that the risk of re-identification to be insignificant, the CJEU did not take the strict zero-sum approach of WP 216.

Moreover, in its 'Guidelines 3/2019 on processing of personal data through video devices' of 29 January 2020, the EDPB provided an example of anonymization of a video footage that contradicts the strict approach of WP 216.[128] The said example concerns a data subject request 'for a copy of his or her personal data processed through video surveillance at the entrance of a shopping mall with 30 000 visitors per day'. According to the EDPB,

> [i]f the controller still processes the material a copy of the video footage should be provided. If other data subjects can be identified in the same material then that part of the material should be anonymised (for example by blurring the copy or parts thereof) before giving the copy to the data subject that filed the request.[129]

It appears that the fact that the original video footage remains in the possession of the controller who collected the personal data does not prevent a copy of such video footage in which some or part of the material was blurred to be deemed to be anonymized in the hands of the data subject that filed the request. The EDPB seems to have ignored the linkability criterion of WP 216 in the example in question.

It is therefore very welcome to learn that the EDPB is planning to reinforce the application of fundamental data protection principles and establish common positions and guidance through guidelines on anonymization and pseudonymization.[130] In our view, the EDPB should look specifically at the challenges of de-identifying and anonymizing unstructured data besides revising WP 216. Among the issues to be clarified are: (i) whether what is stated in WP 216 regarding structured data also applies to unstructured data; (ii) whether one should apply the strict or risk-based approach.

To be clear, we believe that the EDPB should abandon the strict approach to interpretation, which is neither realistic in terms of achieving the aim of anonymization of personal data nor practical and useful given that the anonymization under the strict approach is equal to a complete destruction of data. The risk-based approach mirrors the logic of the GDPR being a risk-based legal framework that does not require that the processing of the personal data should be at zero-risk. Anonymization is a type of processing of personal data and applying a zero-risk approach to only this particular processing is neither logical nor supported by the spirit of the GDPR. The risk-based approach provides more flexibility for the controller in question, but by no means reduces its responsibilities and obligations with respect to the protection of personal data that controller processes. The principle of accountability applies to all types of processing of personal data and is one of the cornerstones of the GDPR.

We believe that when providing guidelines for anonymization and/or constructing a risk-based test, the EDPB should balance the need for concrete, clear, and precise recommendations and the necessity of exercising some margin of discretion by the controller in applying those recommendations. A too detailed test may, in a worst-case scenario, result in attempts to circumvent it. An unclear and too theoretical test may be misapplied or not applied at all. Also, the test must take into account the existing technology—for both anonymization and identification—and its future developments, so as it does not become outdated and thus useless within a short period of time. Importantly, the technology must also be considered in connection to the growing amount of data about data subjects. The test being technology neutral is a matter of course given that the GDPR is technology neutral.[131] We also believe that the EDPB should consider the 'motivated intruder' and 'data environment' as elements of the said test. They mirror the risk based-approach we strongly argue for.

Given the importance of carrying out the process of anonymization in a correct manner and the consequences of falling within or outside the scope of the GDPR, the EDPB should provide examples of how the controller should proceed in different situations and when different types of data (structured or unstructured, text or images etc.) are involved. In this respect, we refer to the EDPB's 'Recommendations 01/2020 on measures that supplement transfer tools to ensure compliance with the

128 EDPB, 'Guidelines 3/2019 on processing of personal data through video devices', Version 2.0, adopted on 29 January 2020.

129 Ibid para 97.

130 EDPB, *EDPB Work Programme 2021/2022* (2021) <www.edpb.europa.eu/system/files/2021-03/edpb_workprogramme_2021-2022_en.pdf> accessed 7 June 2021, 4.

131 As stated in Recital 15: 'in order to prevent creating a serious risk of circumvention, the protection of natural persons should be technologically neutral and should not depend on the techniques used. The protection of natural persons should apply to the processing of personal data by automated means, as well as to manual processing, if the personal data are contained or are intended to be contained in a filing system. Files or sets of files, as well as their cover pages, which are not structured according to specific criteria should not fall within the scope of this Regulation.'

EU level of protection of personal data' where Annex 2 provides useful examples of supplementary measures that may now be required to implement when personal data are transferred to third countries.[132] In the case of anonymization, the abovementioned notions of 'data environment' and 'motivated intruder' require particular attention.

Conclusion

Structured data has captured and dominated the attention of legal scholars, computer scientists, and regulators when addressing the question of anonymization. However, tables, graphs, and other structured data only form a small part of the information that is being used for machine learning and other big data applications. This article has thus sought to move beyond this tip of the iceberg data to understand how privacy requirements in the GDPR affect unstructured data such as text documents or images.

In the absence of a definition of the term 'anonymous data' in the GDPR, we examined its antithesis—personal data—and the identifiability test in Recital 26 GDPR to understand what conditions must be in place for the anonymization of unstructured data. In doing so, we analysed and applied two contrasting approaches for determining identifiability that are prevalent today. The first was the risk-based approach and the second was the strict approach in the Article 29 Working Party's Opinion on Anonymization Techniques (WP 216).

In our view, both approaches accord with a textual reading of the GDPR. However, we argue that in light of the CJEU's developing jurisprudence (especially the *Breyer* case) and the use of other legal methods commonly employed by the CJEU—context and teleology—when literal approaches are inconclusive, a risk-based approach is clearly preferable. This becomes especially clear given the teleological preference for interpretation that does not lead to draconian and absurd consequences. Through two case studies, we illustrated the challenges encountered when trying to anonymize unstructured datasets according to the strict approach. We show that, while the risk-based approach offers a more nuanced test consistent with the purposes of the GDPR, the strict approach of WP 216 makes anonymization of unstructured data virtually impossible as long as the original data continues to exist.

To be sure, we are clear that a risk-based approach is far from a 'free for all'. In many cases, the application of this test can result in requirements for anonymization that are very strict. We underline this contextual variance in application in our discussion of temporality: improving computational methods of re-identification can make the application of a risk-based approach stricter over time if there is not a corresponding improvement in technologies to resist re-identification. Nonetheless, in our view, a risk-approach remains a preferred interpretation that better balances the different objectives of the GDPR and ensures internal coherence in the interpretation of its provisions. A risk-based approach that takes account of the context, environment and threats surrounding the data[133] is more in line with the risk-based approach of the GDPR and CJEU jurisprudence.

At the time of writing, the EDPB has embraced seemingly both the risk-based and strict approach, providing little clarity to the situation. However, it is positive that the EDPB has signalled that it is planning to reinforce the application of fundamental data protection principles and establish common positions and guidance through guidelines on anonymization and pseudonymization. We hope it will abandon the strict approach as well as engage with the question of unstructured data. In particular, the EDPB should provide examples of how the controller should proceed in different situations and when different types of data (structured or unstructured, text or images etc.) are involved. If the EDPB opts to retain the strict approach in respect of structured data, as well as to apply it to unstructured data, then perhaps it is time to acknowledge that EU data protection legislation has really become 'the law of everything' and redirect its focus on providing guidance on different risk profiles rather than on anonymization techniques.[134]

Acknowledgement

We would like to thank Luca Tosoni for his comments on an earlier version of this article. We would also like to thank the following research assistants at the Faculty of Law, University of Oslo, for carrying out the anonymization exercise for the first case study: Isak Falch Alsos, Saba Abadhar, Sigurd Teofanovic, Vilde Katrin Lervik, Sarah Kristin Geisler, Louise Øverås Nilsen, Marlena Zaczek, Ole Martin Moen, Nina Stærnes, Rose Monrad, Selina Ovat and Alexandra Kleinitz Schultz.

132 See <https://edpb.europa.eu/system/files/2021-06/edpb_recommendations_202001vo.2.0_supplementarymeasurestransferstools_en.pdf> accessed 5 August 2021.

133 See, in particular, the section 'The risk-based approach'.

134 See Purtova (n 28). See also Mike Hintze, 'Viewing the GDPR through a De-identification Lens: A Tool for Compliance, Clarification, and Consistency' (2017) 8(1) International Data Privacy Law 86.

This research was performed with the financial support of the CLEANUP (Machine Learning for the Anonymization of Unstructured Personal Data) project funded by the Norwegian Research Council (project number 308904) and CELL project funded by Diku (Norwegian Agency for International Cooperation and Quality Enhancement in Higher Education) (project number 10037). No ethical approval was needed for the conduct of this research and there is no known conflict of interest.

https://doi.org/10.1093/idpl/ipac008