Viewing the GDPR Through a De-Identification Lens:
A Tool for Clarification and Compliance

Mike Hintze[1]

In May 2018, the General Data Protection Regulation (GDPR) will become enforceable as the basis for data protection law in the European Economic Area (EEA). The GDPR builds upon many existing concepts in European data protection law and creates new rights for data subjects. The result is new and heightened compliance obligations for organizations handling data. In many cases, however, how those obligations will be interpreted and applied remains unclear.

De-identification techniques provide a range of useful tools to help protect individual privacy. There are many different de-identification techniques which represent a broad spectrum – from relatively weak techniques that can reduce privacy risks to a modest degree, to very strong techniques that can effectively eliminate most or all privacy risk. In general, the stronger the de-identification, the greater the loss of data utility and value. Therefore, different levels of de-identification may be appropriate or ideal in different scenarios, depending on the purposes of the data processing.

While there is disagreement on certain aspects of de-identification and the degree to which it should be relied upon in particular circumstances, there is no doubt that de-identification techniques, properly applied, can reduce privacy risks and help protect data subjects' rights. Regulatory guidance and enforcement activity under the GDPR can further these key objectives by encouraging and rewarding the appropriate use of de-identification.

Guidance that fully recognizes the appropriate roles of de-identification can also help bring greater clarity to many GDPR requirements. With such guidance, de-identification can become a more practical and useful tool for compliance. But achieving these goals requires an explicit recognition that there is a wide spectrum of de-identification, and that different levels of de-identification have different regulatory and policy implications.

This article examines a number of obligations under the GDPR, including notice, consent, data subject rights to access or delete personal data, data retention limitations, and data security. In each case, it describes how the use of different levels of de-identification can impact the application and interpretation of the requirements and resulting compliance obligations. It proposes that the GDPR requirements in each area should be interpreted and enforced in a way that will encourage the highest practical level of de-identification and that doing so will advance the purposes of the regulation.

European Regulatory Approaches to De-Identification

To date, European data protection law based on the 1995 Data Protection Directive and the regulators' interpretation of it have taken a largely binary approach to de-identification. Data is either personal data and therefore subject to data protection law, or it is anonymous and therefore not subject to data protection law. The Article 29 Working Party's Opinion 05/2014 on Anonymisation Techniques makes

[1] Partner, Hintze Law PLLC. Part-time Professor, University of Washington School of Law. Formerly, Chief Privacy Counsel, Microsoft Corporation. The views expressed in this article are my own and do not necessarily reflect the positions of any current or former employer or client.

clear that the bar for achieving data anonymization is very high. Anonymization must be "irreversible" and the data must be retained in a form in which identification of a data subject "is no longer possible." This state represents the far end of the de-identification spectrum.

However, this binary approach can lead to suboptimal results. For example, an organization that uses personal data for a purpose that cannot be accomplished with fully anonymized data may have insufficient incentive to apply any level of de-identification to the data. The data therefore may be keep in a fully identified state – even if some level of de-identification would be compatible with the purposes and could provide meaningful privacy protections for the individuals. Thus, the binary approach to de-identification can result in levels of de-identification that are lower, and therefore less protective of individual privacy, than they could and should be.

De-Identification Under the GDPR

As with the 1995 Directive, the GDPR recognizes the concepts of both personal data and anonymous data. But compared to the largely binary approach under current European data protection law, the GDPR provides the basis to recognize a much more complete spectrum of de-identification.

The GDPR helpfully adds an explicit recognition of an intermediate level of de-identification with the concept of pseudonymous data. Pseudonymous data is personal data that cannot be attributed to a specific individual without the use of additional information (which must be kept separate and subject to technical and organizational safeguards).

Further, implicit in Article 11 of the GDPR is another level of de-identification. With "Article 11 De-Identified" data, the data controller is "not in a position to identify the data subject." As set out in Articles 11(2) and 12(2), this level of de-identification has significant implications for data controllers' obligations under other articles of the GDPR.

Finally, the definition of "personal data" – which mirrors the definition under the 1995 Directive – provides the basis for yet another important distinction. Specifically, personal data is defined as "any information relating to an identified or identifiable natural person." Unfortunately, little has been made of the distinction between "identified" and "identifiable" and the terms have been treated as effectively equivalent. However, there are important differences. If the person is "identified" by the personal data, that identified data cannot be thought of as de-identified at all. But data in which the person is not identified, but is rather merely identifiable *does* represent a level of de-identification. And while this level includes a range of techniques, including pseudonymization, that may not be as strong as Article 11 De-Identification, it can provide meaningful protection and risk reduction in many circumstances. Thus, guidance under the GDPR should recognize and encourage methods that convert identified personal data into identifiable personal data.

Levels of De-Identification: Terminology and Taxonomy

Meaningful discussions on de-identification require a common taxonomy and set of terms. Terms used to describe different levels of identifiability are used (and misused) in many ways. While efforts to describe and define a full spectrum of de-identification are needed to bring greater clarity to this area,[2]

[2] The work of the Future of Privacy Forum (FPF) on de-identification is important and highly valuable in this regard. It correctly recognizes there are multiple levels of identifiability, and creates a spectrum of

this paper adopts a simplified grouping of four levels of identifiability – focusing on key distinctions that are explicit or implicit in the GDPR as discussed above, and that are the most important for the policy discussions below. It describes four levels of identifiability, referred to as: (1) Identified, (2) Identifiable, (3) Article 11 De-Identified, and (4) Anonymous / Aggregated. Each of the four levels is described below.

Identified data identifies or is directly linked to data that identifies a specific natural person (such as a name, e-mail address, or government-issued ID number).

Identifiable data relates to a specific person whose identity is not apparent from the data; the data is not directly linked with data that identifies the person; but there is a known, systematic way to reliably create or re-create a link with identifying data. Pseudonymous data as defined in the GDPR is a subset of Identifiable data.

Article 11 De-Identified data may relate to a specific person whose identity is not apparent from the data; and the data is not directly linked with data that identifies the person. The data could potentially be re-identified if matched to additional identifying data provided by the data subject, but there is no known, systematic way for the controller to reliably create or re-create a link with identifying data. This data may be subject to potential re-identification attacks that could create a possibility of associating *some* number of records to an identifiable individual with *some* degree of confidence. This category includes data sets that in the past were incorrectly characterized as anonymous and publicly released, such as the well-known cases of AOL search data and the Netflix Prize data, and where some small number of records from the data sets were unexpectedly re-identified.

Anonymous / Aggregate data is (1) stored without any identifiers or other data that could identify the individual or device to whom the data relates; and (2) aggregated with data about enough individuals such that it does not contain individual-level entries or events linkable to a specific person. Anonymization methods must be irreversible and eliminate any known or foreseeable possibility of linking any of the data to an individual to whom the data originally related.

These four levels may be summarized as follows:

	Identified	**Identifiable**	**Article 11 De-Identified**	**Anonymous / Aggregate**
Directly linked to identifying data	Yes	No	No	No
Known, systematic way to (re)identify	Yes	Yes	No	No
Relates to a specific person	Yes	Yes	Yes	No

Each greater level of de-identification provides more protection and further reduces risk to individuals. The first three levels all are personal data within the scope of European data protection law, including

ten levels with meaningful distinctions between each. It will advance the dialogue and help form the basis for greater a consensus on levels of de-identification and terminology. This paper focuses on a smaller number of de-identification levels to simplify the discussion and present the policy proposals at a general level. But the policy arguments and recommendations made in this paper can also be applied directly to the FPF taxonomy, and doing so in the future may be helpful in formulating more detailed and actionable guidance.

the GDPR. Only Anonymous / Aggregate data is completely outside the scope of European data protection law.

GDPR Obligations Viewed Through the De-Identification Lens

The rights and obligations included in the GDPR are more extensive than those under current European data protection law. In the months leading up to the effective date for the GDPR, data controllers and processors are looking for clarity and practical compliance tools. As described below, for many GDPR obligations, de-identification can provide both.

Legal Basis for Processing: Consent or Legitimate Interests

GDPR Article 6 sets out the various bases for lawful processing of personal data. The first basis listed is the consent of the data subject. However, as compared to current law, the GDPR arguably makes it more difficult to obtain and rely on consent. The definition of consent is stricter – requiring that consent be "freely given, specific, informed and unambiguous,"[3] and Article 7 sets out additional requirements a controller must meet to rely on consent. Further, the GDPR is making consent more difficult at a time when technological advances such as the Internet of Things, "big data" analytics, and machine learning are making reliance on consent increasingly impractical in many instances.

When obtaining consent is impractical or impossible, a common (and often only available) alternate basis for lawful processing is the "legitimate interests" of the data controller or a third party. However, this too may be difficult to rely on under the GDPR, creating a dilemma for data controllers. Regulators can provide both clarity and flexibility, while helping to encourage productive uses of data in a way that protects privacy, by providing guidance that reliance on legitimate interests will be looked upon more favorably if the data is de-identified. The greater the degree of de-identification, the easier it should be to rely on legitimate interests for the processing of such data. For instance, controllers should always be able to rely on legitimate interests for the processing of Article 11 De-Identified data. And even lesser degrees of de-identification (such as with Identifiable – including pseudonymous – data) should strengthen the case for relying on legitimate interests.

Article 6(4) of the GDPR supports the idea that de-identification can be used to help justify a basis for lawful processing other than consent. "Where the processing for a purpose other than that for which the personal data have been collected is not based on the data subject's consent . . . the controller shall, in order to ascertain whether processing for another purpose is compatible with the purpose for which the personal data are initially collected, take into account, inter alia . . . (e) the existence of appropriate safeguards, which may include encryption or pseudonymisation."

This approach also is supported by the history of the GDPR and key European officials involved in its development. See, for example, the December 2013 comments of Neelie Kroes, Vice-President of the European Commission responsible for the Digital Agenda:

> Sometimes, full anonymisation means losing important information, so you can no longer make the links between data. That could make the difference between progress or paralysis. But using

[3] By comparison, the definition on the 1995 Data Protection Directive required consent to be "freely given, specific, and informed." The addition of "unambiguous" could be read as raising the bar on what may constitute valid consent.

> pseudonyms can let you to analyse large amounts of data: to spot, for example, that people with genetic pattern X also respond well to therapy Y. So it is understandable why the European Parliament has proposed a more flexible data protection regime for this type of data. Companies would be able to process the data on grounds of legitimate interest, rather than consent. That could make all the positive difference to big data: without endangering privacy.[4]

While the final draft of the GDPR backed off from a blanket rule that pseudonymous data automatically qualifies for processing on the basis of legitimate interests, the regulation is consistent with an interpretation that the stronger Article 11 De-Identification should qualify for legitimate interests, and lesser levels of de-identification that make data merely Identifiable (including pseudonymization) create a strong case for reliance on legitimate interests. The result of such guidance will be that when a data controller wishes to (or needs to) rely on legitimate interests as a basis for processing data, the controller will de-identify the data to the maximum extent compatible with the purposes of the processing in order to strengthen its legal position with respect to its legitimate interests.

Notice

A longstanding issue in data protection law is how prominently a notice must be provided. In many cases, organizations rely on discoverable notice, such as a description of a data practice in a privacy statement. In other cases, regulators have insisted that notice of certain data processing be more prominent. But where and how to make those distinctions is often unclear.

Additional clarity can be provided under the GDPR by making clear that the level of de-identification can play a large role in determining the appropriate prominence of the notice. The more strongly de-identified the data is, the more likely discoverable notice will be appropriate. Particularly for Article 11 De-Identified data, discoverable notice should almost always be sufficient. For identified or identifiable data, discoverable notice may be appropriate, but other factors such as the sensitivity of the data and the expected use will also play a role in determining the appropriate prominence of the notice.

Data Retention

Article 5(e) of the GDPR establishes the general rule that personal data may be "kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed." While implicitly recognizing the value of de-identification, this provision is subject to different interpretations.

It is clear that full anonymization is an alternative to deletion when the data is no longer "necessary." It is possible, but less clear, that Article 11 De-Identification would suffice. However, the "necessity" standard in this provision also creates uncertainty. Often, data is very, very useful for the purposes for which it is processed. The data may make the processing much more effective, efficient, or accurate; but one could argue both sides of the proposition that the retention of that data is strictly necessary. Such uncertainty could be mitigated in many cases by guidance that de-identification, which lowers privacy risks with respect to that data, should give controllers more flexibility to retain the data for a

[4] Neelie Kroes, "Data isn't a four-letter word," IAPP Europe Data Protection Congress/Brussels, 11 Dec. 2013. Available at http://europa.eu/rapid/press-release_SPEECH-13-1059_en.htm.

longer period. And such guidance will, again, provide a strong incentive to apply the strongest level of de-identification compatible with the purposes of processing.

Data Security

Article 32 of the GDPR requires organizations to implement security measures sufficient "to ensure a level of security appropriate to the risk." The text calls out the risks resulting from "accidental or unlawful destruction, loss, alteration, unauthorised disclosure of, or access to personal data" as particular factors in determining the appropriate level of security.

There are several considerations that are important in determining those risks, such as the nature and sensitivity of the data. But those risks are all significantly reduced when de-identification is applied. And the stronger the level of de-identification, the lower those risks become. Thus, when strong de-identification is applied to data, relatively modest security measures should suffice. Data that is protected with weaker de-identification will require more robust security measures. And data that has not been de-identified at all will require even stronger security.

The text of the GDPR, as well as a prior opinion of the Article 29 Working Party, provides support for this approach, albeit in an indirect way. In both the Working Party opinion on Anonymization referenced above and in certain provisions of the GDPR, pseudonymization is characterized as a security measure (rather than characterizing both pseudonymization and anonymization as points on the de-identification spectrum).[5] But the implication of that view is that by employing pseudonymization (and presumably other de-identification mechanisms), the need for "other" security measures is reduced because the totality of measures taken to protect the data is enhanced.

Data Subject Rights of Access, Deletion, and other Controls

Article 12(2) of the GDPR specifies that if the controller can demonstrate that it is not in a position to identify the data subject (i.e., Article 11 De-Identified data), it need not comply with Articles 15 to 22. Those articles include the right of access (Article 15), rectification (Article 16), erasure (Article 17), data portability (Art. 20), and the right to object to the processing of personal data or obtain a restriction of such processing under certain circumstances (Articles 18 and 21).

This provision reflects the reality that a data controller simply cannot offer these types of user rights and controls if the controller has employed a level of de-identification that precludes it from reliably linking the data back to the individual seeing to exercise these rights. It recognizes that the enormous privacy benefits of encouraging strong de-identification outweigh any inability of a data subject to exercise certain rights, and it reflects the beneficial Article 11 rule that "[i]f the purposes for which a controller processes personal data do not or do no longer require the identification of a data subject by the controller, the controller shall not be obliged to maintain, acquire or process additional information in order to identify the data subject for the sole purpose of complying with this Regulation."

This same distinction will apply to virtually any user control that relates to previously-collected data. If there is no systematic and reliable way to connect the data to the individual, it is impractical or impossible to give the individual control over that data.

[5] See the Article 29 Working Party's Opinion 05/2014 on Anonymisation Techniques; see also the reference in GDPR Article 6(4) to "appropriate safeguards, which may include encryption or pseudonymization."

The key privacy issues discussed above are represented in the following table.

<table>
<tr><th></th><th>Consent or Legitimate Interests</th><th>Notice to Data Subjects</th><th>Data Retention Limitations</th><th>Appropriate Data Security</th><th>Access, Erasure, Controls</th></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>Identified</td><td rowspan="3">Consent of Data Subject
↕
Legitimate Interests</td><td rowspan="3">Prominent Notice
↕
Discoverable Notice</td><td rowspan="3">Shorter Retention
↕
Longer Retention</td><td rowspan="3">Stronger Protections
↕
Some Protections</td><td rowspan="2">Required</td></tr>
<tr><td>Identifiable</td></tr>
<tr><td>Article 11 De-Identified</td><td rowspan="2">No Requirement</td></tr>
<tr><td>Anonymous / Aggregated</td><td colspan="4">No Requirements</td></tr>
</table>

Other Privacy Benefits

The usefulness of the de-identification lens is not limited to the issues discussed above. Many other obligations under the GDPR could be clarified if viewed through a de-identification lens, and guidance that encourages the use of de-identification will have many more privacy benefits than those already specifically noted.

For example, data breach notification obligations under Articles 33 and 34 of the GDPR are tied to the likelihood of risk to the rights and freedoms of natural persons. The level of de-identification applied to the data has a direct bearing on that risk. The stronger the de-identification, the less likely the risk and the more likely controllers and supervisory authorities should conclude that notification is not warranted. Such an approach will help avoid over-notification and ensure notifications are focused on those instances where there is a real risk to individuals.

Another important example involves government access to personal data held by private sector organizations. This issue is at the heart of cross-border data transfers, raises important concerns regarding individuals' privacy and other fundamental rights, and often dominates privacy discussions. Guidance encouraging the use of de-identification can help mitigate those concerns. Fully anonymized data cannot be tied to an individual person. And in many cases, companies could refuse or resist government demands for de-identified data, and especially Article 11 De-Identified data, due to the inability to reliably and systematically connect the data to an identified subject of an investigation.

Finally, a recognition that de-identification involves a wide spectrum of practices with different levels of strength, along with clear guidance tied to those levels that provides greater regulatory relief for more strongly de-identified data, can help remove the anxiety and hyperbole that dominate many discussions about the scope of "personal data." Too often, discussions about whether a IP address or other unique identifier meets the definition of "personal data" are characterized as an all-or nothing debate. Either the data is subject to the full range of obligations under data protection law, or it is subject to none. This conclusion is incorrect (even under current law), but clearer and more explicit recognition of the de-identification spectrum can change the nature of the debate to a more nuanced and productive discussion of what obligations should apply and how, depending on the nature and identifiability of the data.

Conclusion

This article largely focuses on the GDPR requirements. But the same analysis and same arguments can apply to other privacy laws and can be used by privacy regulators around the world in interpreting and applying those laws.

Recognizing that there is a broad spectrum of de-identification, and identifying certain key points along that spectrum, has important regulatory and policy implications. It enables the development of regulatory guidance that encourages the maximum use of de-identification compatible with the purposes of the data processing. That, in turn, can provide the optimal balance between maintaining utility of data and protecting the privacy of individual data subjects. Such guidance can also help provided much-needed clarity related to new GDPR obligations. In sum, viewing the GDPR through the de-identification lens can be a win-win-win for regulators, data controllers, and individual data subjects alike.