Skatteforsk
Centre for Tax Research

February 2025 / Note 05-2025

The Difference Between Personal Data and Anonymous Data

Co-authored by:

Anniken M. Henriksen & Hanne Pernille Gulbrandsen
Deloitte Advokatfirma AS

Hector Ulloa
Skatteforsk – Centre for Tax Research

Skatteforsk
Centre for Tax Research

Skatteforsk – Centre for Tax Research is an independent research center based at the School of Economics and Business at the Norwegian University of Life Sciences (NMBU). Our mission is to bridge the gap between cutting-edge academic research and practical tax policy. A key aspect of our work is the use of pseudonymized and anonymized data, which plays a vital role in driving innovation and advancing the academic frontier.

In 2024, Skatteforsk has focused extensively on expanding our expertise in data usage for research purposes. Highlights include hosting a special session on the “Use of Leaked Data in Academic Research” at the 80th Annual Congress of the International Institute of Public Finance in Prague and organizing a workshop series with Norwegian public institutions on “The Public Sector’s Use of Its Own Data.” The outcomes of these initiatives will be published later as part of this Notes Series.

Given the importance of clearly defining what we mean by anonymized/pseudonymized data and anonymization across our work, this note lays the groundwork for these discussions. It serves as a foundational building block for outcomes from initiatives like those mentioned above and future efforts in this area.

For more publications visit: https://skatteforsk.no/

Contact us at: skatteforsk@nmbu.no

Skatteforsk’s Notes Series is financed through support from The Research Council of Norway (Grant number 341289 & 352151).

Contents

1) Introduction

The General Data Protection Regulation (GDPR) governs the processing of personal data. When data is considered anonymous, it falls outside the scope of GDPR. In summary, the concept of anonymous data under GDPR is a dynamic and evolving area, with the European Court of Justice (ECJ) and the European Data Protection Board (EDPB) providing essential guidance and clarifications. The key principle is that data can be considered anonymous if it is not related to an identified or identifiable person, even if theoretical re-identification is possible as long as it is not practically achievable without disproportionate efforts. Controllers and processors must carefully assess and apply anonymization techniques to ensure compliance with GDPR's data protection principles.

Personal data may be made anonymous through a process where the possibility of identifying individuals in the actual data set is removed. Even though one does not have access to direct identifiers, a data set may be regarded as personal data while identification may be possible through use of technical means and open data, i.e., data from social media, official registries etc.

The assessment of which information is considered personal data, pseudonymized data or anonymous data is important as the potential non-compliance with the GDPR can have major consequences, both financially and in terms of reputation. Therefore, it is important to keep this issue in mind, and carefully assess if there is a risk of re-identification of information in data sets and/or to be used for scientific purposes.

2) Personal Data

Personal data encompasses any information that can be linked to an identified or identifiable natural person. This definition is typically interpreted broadly. If a natural person can be identified, directly or indirectly, through factors like a name, an identification number, location data, an online identity, and more, the data is considered personal data. Aspects related to a person's physical, physiological, genetic, mental, economic, cultural, or social identity can also be regarded as personal data if they contribute to identification and always when linked to a direct identifier. Please note that there is no requirement that the information is true or proven, meaning that untrue information about an identifiable person is still considered personal data.[1] Both objective and subjective information is considered personal data.[2]

It derives from GDPR recital 26 "*that 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****".*

In the data protection directive from 1995 you find a similar outset; thus, this term has been subject to interpretation and development through several court cases in the European Court of

[1] Note that the GDPR has rules that envisages the possibility that information is incorrect and provide for a right of the data subject to access that information and the right to rectification of inaccurate personal data.

[2] Article 29 Data Protection Working Party, Opinion 4/2007 on the concept of personal data, page 6. Please note that the Article 29 Data Protection Working Party was an independent European advisory body on data protection and privacy and can be considered as the predecessor of the European Data Protection Board (EDPB).

Justice (ECJ). To understand the concept of anonymous data, it is relevant to have some basis knowledge about this caselaw.

Some personal data are considered more sensitive than other personal data such as e.g., name, and e-mail, so-called **special categories of personal data**. Due to the sensitiveness, and the potential significant risks to the fundamental rights and freedoms, the GDPR sets out additional requirements when processing such types of personal data.[3] Special categories of personal data are exhaustive defined in the GDPR article 9 (1) as personal data revealing e.g., political opinions, religious or philosophical beliefs, or trade union membership, as well as the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation. Processing of such personal data are as a main rule prohibited, but exceptions can be made if one of the alternatives in the GDPR article 9 (2) letters a-j applies, e.g.:

- the data subject has given explicit consent to the processing cf. letter a,
- the processing is necessary for the purpose of carrying out the obligations exercising specific rights of the controller or of the data subject in the field of employment and social security cf. letter b, or,
- the processing is necessary for reasons of substantial public interest, on the basis of Union or Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject, cf. letter g.

The use of national identity number and other unique identifiers, e.g., biometric means of identification, are specifically regulated in the Norwegian Privacy Act (IN: *personopplysningsloven*) section 12 cf. the GDPR article 87. These types of personal data are not considered special categories of data, cf. GDPR article 9 (1), however they are still considered more worthy of protection, meaning, processing of such personal data requires a greater degree of care. Therefore, an additional requirement is set out in section 12:

> National identity number and other unique identifiers can only be processed when there is a *legitimate need* for definite identification and the method is *necessary* to obtain such identification. A *legitimate need* must be assumed to be more than mere convenience.

As an example, in many public registers that are used to assign rights and obligations, there is a need for definite identification. The criterion of necessity is considered fulfilled if other, less secure means of identification, are not sufficient, e.g., a customer number, name, or address.[4] It is also of relevance how important secure identification is for the data subject in question, for instance what consequences an identity confusion could lead to.[5]

3) Pseudonymized Data

Pseudonymized data must not be confused with anonymous data. Pseudonymized data is still considered personal data, but it can no longer be attributed to a specific individual without the use of additional information, like an identification key. To qualify as pseudonymized data, the

[3] The GDPR article 9 (1) and (2) as well as GDPR recital 51.

[4] Ot.prp. nr. 92 (1998-1999), comments to section 12, page 114. These preambles are relevant because section 12 in the Norwegian Privacy Act of 2018 continues applicable law according to the previous Norwegian Privacy Act of 2000.

[5] Ot.prp. nr. 92 (1998-1999), comments to section 12, page 114

identification key must be kept separate and be subject to measures ensuring that the personal data cannot be linked to the data subject.[6] Pseudonymized data is subject to GDPR but often with more lenient regulations. The application of pseudonymization to personal data can reduce the risks to the data subjects concerned and help controllers and processors to meet their data-protection obligations.[7]

4) Anonymous Data

A. What Is Anonymous Data?

Anonymous data has no relation to an identified or identifiable person.[8] The processing of anonymous data is not regulated by GDPR. Consequently, data considered anonymous can be collected, recorded, transferred, and stored without limitations.

When personal data is rendered anonymous to the extent that the data subject is no longer identifiable, it becomes equivalent to anonymous data.

B. The Concept Of Anonymous Data Is A Complex Issue

According to GDPR, an anonymized dataset is one where all personally identifiable information is permanently removed. Another way to look at the concept of anonymous data, is that data is only anonymous if it is not possible to revert to personal data. However, data can be de-anonymized through various methods, and the determination of identifiability should consider all possible means likely to be used. This classification's impact on GDPR applicability should be recognized, as the "means likely to be used" will evolve with changing technology.

The assessment of whether the data allow identification of an individual, and whether the information can be considered as anonymous or not depends on the circumstances, and a case-by-case analysis should be carried out with particular reference to the extent that the means are likely reasonably to be used for identification.[9] Additionally, as laid down in GDPR recital 26, the effort, time, and cost required for de-anonymization must be considered. As stated by the Article 20 Data Protection Working Party, this is particularly relevant in the case of statistical information, where despite the fact that the information may be presented as aggregated data, the original sample is not sufficiently large and other pieces of information may enable the identification of individuals.[10]

C. The Motivated Intruder Test

In 2015, the Norwegian Data Protection Authority (the Authority) issued a guide on various anonymization techniques.[11] The guide was updated in 2019, and it is still considered valuable. The guidelines largely refer to the Article 29 Group's Opinion 05/2014 on Anonymization Techniques.[12] In this guideline WP29 states the difference between pseudonymized and anonymized data, as now defined in the GDPR. The legal basis for the guide is recital 26 and case law related to the sentence "... account should be taken of all the means reasonably likely

[6] GDPR article 4 (5)
[7] GDPR recital 28.
[8] Article 29 Data Protection Working Party, Opinion 4/2007 on the concept of personal data, page 21.
[9] Ibid.
[10] Ibid.
[11] The guide from the Data Protection Authority can be found her: anonymisering-veileder-041115.pdf (in Norwegian only, but entails description of some anonymization Techniques and their pros and cons)
[12] Article 29 Data Protection Working Party, opinion 05/2014 on Anonymisation Techniques

to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly".

The Authority defines anonymization as a process of removing the possibility of identifying individuals in a data set. The anonymization technique must be determined in each case, based on the characteristics of the data set in question. It must be noted that anonymization only exists if the process is irreversible.

Independently of the techniques and methods used, anonymized data sets inherently carry a risk of re-identification, meaning, that one manages to identify individuals from **initially** assumed anonymous data sets.

The framework given to us in recital 26 clarifies then that through anonymization, **it should not be possible to find the link between the anonymized information and an individual**, when taking into account all means that, to the best of one's knowledge, **could reasonably be used** to identify an individual from that data.

In today's society, the risk of re-identification is great due to enormous access to publicly available data and powerful analytics technology. The Authority points out that studies have shown that by collating data from several sources, it is possible to re-identify people by only knowing two attributes in an anonymized data set, such as postcode and date of birth. Re-identification can occur by someone taking personal data they already have about others and searching for matches in an anonymized data set, or by taking a match from an anonymous data set and searching for further matches on publicly available information such as information from public registers or social media.

The Authority notes that the most important method for ensuring that anonymization is sufficient is to carry out a so-called re-identification test, also called a motivated intruder test.[13] In this test, a motivated intruder should be assessed as a person/organization that without prior knowledge tries to identify individuals in an anonymized data set. The test assesses whether a motivated intruder, who starts without any prior knowledge but wishes to identify an individual from whose personal data the anonymous information is derived, is likely to be successful. The intruder shall be regarded as sufficiently competent and shall be assumed to have access to resources such as the Internet, public registers and libraries, etc. and that he or she is able to use various investigative techniques to make contact with and access people with knowledge of the identity of individuals in the data set.

Based on GDPR recital 26 and means *reasonably likely to be used*, it appears that the motivated intruder should neither be assessed based on specialist knowledge, expertise in computer hacking, nor need to have access to special equipment to break in to access data in secure storages.

D. Case Law From The European Court Of Justice (ECJ)

A summary of case law from ECJ shows that the term personal data is subject to broad interpretation, but if identification requires disproportionate effort and resources, it may be regarded as anonymous to the user in question.

[13] See pages 11-12 in the guide: anonymisering-veileder-041115.pdf

Joined Cases C-141/12 and C-372/12 (2014) Y.S. v. Minister for Immigration, Integration and Asylum

Context:
This case dealt with to access to minutes concerning the temporary residence permit as a right of asylum in the Netherlands.

Findings:
In this case, the ECJ ruled that data can be considered anonymous if the individual's identity is irreversibly removed. The court emphasized that even in cases where there is a theoretical possibility of re-identification, as long as it ***requires disproportionate efforts and resources***, the data may still be regarded as anonymous.

Case C-582/14 Patrick Breyer v. Federal Republic of Germany

Context:
This case addressed the issue of dynamic IP addresses and whether they constitute personal data or anonymous data.[14]

Findings:
The ECJ held that dynamic IP addresses can be considered personal data if the data **controller has the legal means to obtain additional information necessary for identification,** i.e., with additional data held by the internet service provider. However, if the data controller does not possess the means to identify the data subject, then the IP address might be processed as anonymous data.

Case C-434/16 Peter Nowak v Data Protection Commissioner

Context:
The case focused on whether written answers submitted in an examination and the examiner's comments can be classified as personal data.[15]

Findings:
"The CJEU concluded that the written answers of an examinee, linked with the candidate's name or other identifier, constitute personal data, as they reflect the extent of the individual's knowledge and competence". Further of relevance: "Contrary to what the Data Protection Commissioner appears to argue, **it is of no relevance**, in that context, **whether the examiner can or cannot identify the candidate at the time when he/she is correcting and marking the examination script**. For information to be treated as 'personal data' within the meaning of Article 2(a) of Directive 95/46, **there is no requirement that all the information enabling the identification of the data subject must be in the hands of one person** (judgment of 19 October 2016, Breyer, C-582/14, EU:C:2016:779, paragraph 43). It is also undisputed that, in the event that the examiner does not know the identity of the candidate when he/she is marking the answers submitted by that candidate in an examination, the body that set the examination, in this case the CAI, does, however, **have available to it the information needed to enable it easily and infallibly to identify that candidate through his identification number, placed on the examination script or its cover sheet, and thereby to ascribe the answers to that candidate**.

[14] Read the case here: EUR-Lex - 62014CN0582 - EN - EUR-Lex
[15] Read the case here: EUR-Lex - 62016CJ0434 - EN - EUR-Lex

Case C-319/22

Context:
The case examined whether VIN (Vehicle Identification Number) should be considered personal data.[16]

Findings:
The court concluded that VINs do constitute personal data within the meaning of Article 4(1) of the GDPR if the person accessing them has the means to link the VIN to an identified or identifiable natural person. This interpretation is based on the nature of VINs and the possibility of linking them to specific individuals, such as vehicle owners, using other available data like vehicle registration certificates.

E. Disproportionate Effort And Resources

Relevant case law sets out "***disproportionate efforts and resources***" as a criterion when determining "all the means reasonably likely to be used" when it comes to re-identification. What is considered "disproportionate efforts" may change over time due to the constant technological developments, for instance the development of large language models and other special tools. Efforts that were considered disproportionate three years ago, may not be disproportionate today. Therefore, it is not useful to present an exhaustive list of when identification is not possible. A specific assessment must be made. Some key factors can be considered in this regard, including the expense, time, effort, and know-how needed to implement the re-identification means.

In Case C-582/14 *Breyer* para 45 and 46 the criterion is described in more detail:

> However, it must be determined 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.
>
> And further: "…that would not be the case if the identification of the data subject was prohibited by law or **practically impossible** on account of the fact that it requires a **disproportionate effort in terms of time, cost and man-power, so that the risk of identification appears in reality to be insignificant**."

The CJEU concluded in *Breyer* that the German Federal Republic institution had lawful access to 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, and this was considered "means which may likely reasonably be used in order to identify the data subject".

The French Commission Nationale e l'Informatique et des Libertés (CNIL), sets out the same requirement as the CJEU, describing anonymization as processing which consists in an ensemble of techniques which render identification of the data subject "practically impossible.[17]

The bar seems very high when it comes to the risk of re-identification. However, the requirement is not impossibility in general, but *practically* impossibility. The motivated intruder test may be used as a tool to decide whether reidentification is possible in practice.

[16] Read the case here: EUR-Lex - 62022CA0319 - EN - EUR-Lex
[17] The French Commission Nationale e l'Informatique et des Libertés (CNIL): https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles

The increasing development of large language models may affect what is considered practically impossible. Large language models are large deep learning models that are pre-trained on vast amounts of data and can recognize and analyze text as well as other tasks. Such models may be used as a tool for re-identification, as well as other tools. However, the use of such tools may be expensive and require special permits for use, which may affect the question of practical impossibility as well as the question of which the financial situation in a company may affect the conclusion of the assessment.

5) Conclusion

As mentioned in this note's introduction, personal data may be made anonymous through a process where the possibility of identifying individuals in the actual data set is removed. Independently of the techniques and methods used, anonymized data sets inherently carry a risk of re-identification. What is considered a disproportionate effort for re-identification may change over time due to the constant technological developments of our times, for instance the development of large language models and other special tools. Efforts that were considered disproportionate before, may not be disproportionate today.

I must also be noted that even in cases where one does not have access to direct identifiers, a data set may be regarded as personal data if identification is possible through the use of technical means to connect one's information with open data such as public registries and social media.

The assessment of which information is considered personal data, pseudonymized data or anonymous data is important as the potential non-compliance with the GDPR can have major consequences, both financially and in terms of reputation. Therefore, it is important to keep this issue in mind, and carefully assess if there is a risk of re-identification of information in data sets and/or to be used for scientific purposes.

6) Sources and Links for Further Reading

- Article 29 Data Protection Working Part, Opinion 4/2007 on the concept of personal data: 12251/03/EN

- Misunderstandings related to anonymization: https://edps.europa.eu/system/files/2021-04/21-04-27_aepd-edps_anonymisation_en_5.pdf

- Opinion 05/2014 on Anonymisation Techniques: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

- Guidelines from the Norwegian DPA (in Norwegian only) (Contains a description of some anonymization techniques, and their pros and cons): https://www.datatilsynet.no/globalassets/global/dokumenter-pdfer-skjema-ol/regelverk/veiledere/anonymisering-veileder-041115.pdf

- They who must not be identified—distinguishing personal from non-personal data under the GDPR, Michèle Finck, Frank Pallas: https://academic.oup.com/idpl/article/10/1/11/5802594

- CNIL on anonymization: L'anonymisation de données personnelles | CNIL